GESTURE RECOGNITION BY FOURIER ANALYSIS TECHNIQUES

A Thesis submitted for the degree of Doctor of Philosophy

by

Peter Reginald George Harding

School of Computing and Advanced Technologies

Buckinghamshire Chilterns University College,
The City University

June, 2007

Abstract

Recent linguistic research has shown that gesturing is an important channel of non-verbal communications that augments meaning to the spoken word. This thesis demonstrates that a hand gesture can be modelled in 2DT (two dimensions and time) as an aperiodic waveform. Fourier analysis of the waveform generates positive and negative sequence harmonic components from which characteristics of the gestures can be recognised. Variability in gesture lengths are confronted by re-sampling the data to a fixed length, from which harmonic components can be effectively compared. Manipulation of the harmonic data gives the gesture data scale and translation invariant properties. Gesture characterisation is revealed by harmonic 'orientation' angles and by each harmonic having a unique 'elliptical-corkscrew'. The first three harmonics are generally sufficient to characterise a gesture. Gesture recognition is accomplished by using clustering techniques on the low order harmonic data to select target gestures for a Probabilistic Neural Network (PNN). The PNN requires minimal training and in association with clustering techniques, can select target gestures to reveal inter-class and intra-class differences of gestures ensembles. The application of Fourier analysis to gesture stimuli shows their predominantly oscillatory and idiosyncratic nature. A reliable technique for recording hand coordinate data was developed that fused skin-colour and motion cues. As a result, objects, which when rank ordered by area, invariably related the most significant object to the dominant gesturing hand. An object selection algorithm corrected for most tracking mistakes. The technique has been successfully extended to track two hands simultaneously. The gesturing of one person was followed when there were three people in the scene. Additional observations of the individual and personalised nature of hand gesture to gesture stimuli by the gesturer, has revealed the potential of a prime alternative method of a vision-based biometric.

Contents

Abstract 3
Contents 5
List of Illustrations 9
List of Tables 23
Acknowledgements 29
Author's declaration of previously published work 31
Abbreviations 33

1. Introduction 37
1.1. The origin of gesture 38
1.2. Human-Computer Interface Applications 39
1.3. Aims 42
1.4. Thesis content and organisation 43
2. Review of Hand Gesture Analysis and Gesturing Systems 45
2.1. Introduction 45
2.2. Gestural Analysis and Characteristics 46
2.3. Gesture Recognition Techniques 50
2.3.1. Spatial Modelling 51
2.3.2. State-Based Modelling 51
2.3.3. Alternative Modelling Techniques 53
2.4. Ground Truth Data and Comparisons 54
2.5. Summary 56
3. Detecting Hand Position by Colour and Motion 59
3.1. Introduction 59
3.2. Colour Fundamentals and Models 61
3.2.1. CIE Definitions 61
3.2.2. Colour Models used for Skin-Colour Detection 63
3.2.3. Colour Model Comparisons 66
3.3. The HSV colour-model 67

3.4. Colour Space Experiments 69
3.5. Fusing of Motion and Skin Colour Cues 71
3.5.1. Motion Experiments 74
3.5.2. Rank Ordering of Motion Objects 75
3.5.3. Production of Skin-Colour and Motion Objects 77
3.6. Hue and Motion Segmentation Discussions 78
3.6.1. Combination of Colour, Motion and Edge information 83
3.7. Hue and Saturation in other environments 83
3.8. Summary 84
4. Time Domain Tracking and Normalisation 87
4.1. Introduction 87
4.2. Previous approaches to tracking 88
4.3. Data generated from gesture sequences 90
4.4. The complete OSA (Object Selection Algorithm) 96
4.5. Choice of Sequence Parameters 101
4.6. Three people in an image 104
4.7. OSA Performance 106
4.8. Time Normalisation 108
4.8.1. Decimation and Interpolation 109
4.8.2. Ratio calculation 109
4.8.3. Aliasing considerations as a result of normalisation 111
4.9. Summary 111
5. Fourier Analysis of Gesture Trajectory 113
5.1. Fourier Analysis Applications 113
5.2. Fourier Analysis Concepts 115
5.3. Fourier Analysis in 1D, 2D and 2DT domains. 117
5.3.1. One-Dimensional (1D) Fourier Analysis 117
5.3.2. 2D Fourier Analysis – Fourier Descriptor 119
5.3.3. Gesture trajectory analysis (2DT) 119
5.3.4. Developing exponential equations (positive and negative sequence components) to describe gesture trajectories in 2DT space 121
5.3.5. Exponential synthesis of waveforms 124
5.3.6. Cross-over in trajectory contour 126

5.3.7. Translation, Scale and Orientation, θ_k considerations 131
5.4. Interpretation of harmonic content with simulated gestures 132
5.4.1. Planar triangular trajectory 132
5.4.2. Curved and Oscillatory Trajectories 136
5.5. Analysis and synthesis comparison 137
5.6. Analysis of harmonic components from some gestures 139
5.7. Performance Assessment 142
5.7.1. Gesture Truncation 142
5.7.2. OSA Performance Based on Harmonic Analysis 145
5.8. Pointing Gesture Experiments and Initial Interpretation 147
5.9. Summary 150
6. Gesture Recognition using Probabilistic Neural Networks and Hierarchical Cluster Techniques 153
6.1. Introduction 153
6.2. Probabilistic Neural Network 154
6.3. Clustering 156
6.3.1. Distance Metrics 156
6.3.2. Partitioning methods 158
6.3.3. Cluster validation 158
6.4. Recognition of Pointing Gestures 159
6.4.1. Normalisation of frequency data 159
6.4.2. Recognition using the PNN 160
6.5. Recognition using Clustering Techniques 162
6.5.1. Testing clustering techniques 162
6.5.2. Alternative clustering input data 167
6.6. Summary 168
7. Gesture Experiments 171
7.1. Introduction 171
7.2. Repeated gesture (Intra class variations) 173
7.3. The 'Take Mug' Gesture (Inter class variations) 176
7.3.1. Sampling Rate Adjustments 177
7.3.2. Analysis of the 'Take Mug' Gesture Suite 178
7.3.3. Variations of the 'Take Mug' Gesture Suite 180

7.3.4. Clustering of 'Take Mug' Harmonics 181
7.3.4.1. First harmonic orientation angle cluster 181
7.3.4.2. Second and third harmonic clusters 183
7.4. Gesture Stimuli Experiments 190
7.4.1. The Gesture Stimuli Experiments 190
7.4.2. Frequency Analysis of Gesture Stimuli 193
7.4.3. Additional Observations 200
7.5. Summary 201
8. Conclusions and Future work 203
8.1. Summary 203
8.2. Conclusions 206
8.3. Future research areas 208
References 211
Appendix I – Avatars 223
Appendix II – Skin-Colour Variations due to Different Illuminants and White Balance Corrections 229
Appendix III – Sequence Parameters 245
Appendix IV - Multirate Ratios 261
Appendix V– Interpretation of Harmonic Data 263
Appendix VI – Cluster Analysis - Matlab Toolbox (2006) 285
Appendix VII – Four Gesture Experiments 323

List of Illustrations

Figure 2.1 Taxonomy of Hand/Arm Movements (Source Pavlovic et al., 1997) 47
Figure 3.1 CIE Yxy model comparing Pantone, Monitor and SWOP-CMYK colour gamuts (source: Agfa) 62
Figure 3.2 CIE L*a*b* model (Source: Agfa) 63
Figure 3.3 RGB proportions and related Hue values 68
Figure 3.4 The hue hexacone using positive and negative numbers 69
Figure 3.5 The cross-sectional profile of a simple moving object is shown at frames F1 and F2. The thresholded absolute difference picture 'ADP_{12}' is logically ANDed with the Hue-Saturation mask HS_2 , obtained from F_2 to produce the skin-colour and motion mask, SCM. The SCMI mask indicates the original HS mask region congruent with the SCM mask. 73
Figure 3.6 Difference and absolute difference pictures of frames 1 and 2 before thresholding 75
Figure: 3.7 Two different thresholds of absolute different pictures 75
Figure 3.8 Six largest objects as a result of rank order by area filtering 76
Figure 3.9 Hue-Saturation mask (left), gesture objects from SCM mask (right) 77
Figure 3.10 The position of the first three most significant objects shown by the red, green and blue crosses, respectively on the left image and the assigned or SCMI object in the right image (image size 288 x 360) 78
Figure 3.11 The position of the first three most significant objects shown by the red, green and blue crosses, respectively on the left image and the assigned or SCMI object in the right image (image size 144 x 180) 79
Figure 3.12 The second gesture object (green cross) in the left image detecting motion in the face region and shown as an SCMI object in the right image 80
Figure 3.13 Second and third gesture objects detecting background because of poor segmentation. 80
Figure 3.14 Overlapping hue ranges of 'width' of 0.02 and starting at –0.01 (top left image) and finishing at +0.1 (bottom right image), incremented by 0.01 and shown in red. 82

Figure 3.15 Combining the overlapping hue ranges (Fig 3.16) with the motion mask to produce skin-colour and motion mask in red. 82
Figure 4.1 Gesture object shown by red cross (left) and the assigned or SCMI object (right) showing the hand outline. 91
Figure 4.2 A sequence of 51 frames (time index) showing 2D (left) and 2DT (right) images of the positions of the three most significant rank ordered SCM objects (red, green and blue, respectively). 92
Figure 4.3 A sequence of 51 frames (time index) showing 2D (left) and 2DT (right) images of the first three most significant rank ordered SCMI objects (red, green and blue, respectively). 93
Figure 4.4 2D (left) and 2DT (right) images of the first three most significant rank ordered SCME objects (red, green and blue, respectively). 93
Figure 4.5 2D (left) and 2DT (right) images of the first three most significant rank ordered SCMEI objects (red, green and blue, respectively. 94
Figure 4.6 Comparison of the most significant data (red 'o') with manually obtained data (blue '.') and output position (black '+'). 94
Figure 4.7 The inclusion of the second most significant object (green diamond) improves the tracking performance output (black cross) 95
Figure 4.8 Initial coordinates shown by magenta dotted line and the red dotted line showing the upper and lower tolerance to this estimate. Stopping criteria set by the updated initial position as shown by the cyan coloured line. 96
Figure 4.9 Initialisation of the OSA to set initial trajectory coordinates 98
Figure 4.10 OSA using just two gesture objects to follow the dominant hand 99
Figure 4.11 OSA using just two gesture objects to follow both hands 100
Figure 4.12 An image from a sequence recorded in low illumination and poor white balance 101
Figure 4.13 The most significant SCM object for the two hand sequence with the left hand visually obtained data (cyan) and the most significant object coordinates (red) 102
Figure 4.14 The most significant SCMI object for the two hand sequence for the optimum Hue and Saturation range with visually/manually obtained data (left hand, cyan) and the most significant object coordinates (right hand, red). 103

Figure 4.15 The first two SCM objects (red circles and green diamonds, respectively) coordinate identify the trajectory of the right and left hand, with cyan and magenta dotted lines showing the visually obtained left and right hand ground-truth data respectively. 104
Figure 4.16 Three people in a PETS image 104
Figure 4.17 Tracking of PETS data with a '+' from first two SCM object coordinates (red circles and green diamonds, respectively) and comparison with visual data (magenta) 105
Figure 4.18 An example of tracking data to access the OSA's performance. 106
Figure 4.19 Bar charts showing difference between manual and OSA for row (upper) and column (lower) coordinates. The hand span is shown between the red lines at 18 pixels and the search window set at 30 pixels (green lines). 107
Figure 4.20 Search distance insufficient for good tracking 107
Figure 4.21 2D (left) and 2DT (right) representation of a gesture trajectory 108
Figure 4.22 2DT view of the normalization of a 47 sample (blue cross) gesture to 64 samples (red circle) by the ratios given for No. 4 of Table 4.1 110
Figure 5.1 Single- and double-sided spectra of $\cos(\omega t+\phi)$. (Source: Bissell and Chapman, 1995) 118
Figure 5.2 Complex exponential, $f(t) = e^{-j\omega t}$, negative frequency (Source: Kraniauskas,1993) 119
Figure 5.3 Three points A, B and C on an elliptical trajectory in appearance-space sampled at t_1, t_2 and t_3 in the time domain 120
Figure 5.4 Pictures of 4 views of an ellipse sampled in the time domain. The top left picture shows the spatial domain representation; the top-right shows the change of 'y' with time (a sine wave); the bottom-left picture shows the change of 'x' with time (a cosine wave)and the bottom right shows one revolution of the 'elliptical corkscrew', i.e. a 2DT view. 121
Figure 5.5 Comparing matrix 'x' and exponential 'o' equations of an ellipse with A = 1.5 and B = 0.5 or $A_p = 1$ and $A_n = 0.5$, with Orientation angle of 30° and phase shift of 60° 123
Figure 5.6 Rotating positive (red) and negative (black) sequences and the resulting (blue) ellipse, with 'o' indicating starting point and '*' end point. 124

Figure 5.7 Four views of the first, third and fifth harmonics at the same orientation. Rotating positive (red) and negative (black) sequences and the resulting (blue) ellipse. The starting point-and end-point of the time sequence indicated on all pictures as 'o' and '*' (blue) respectively. 125

Figure 5.8 Four views of elliptical structure, for a third harmonic, with A_p at 0.042 and A_n at 0.103, orientation angle of 38° and phase shift of 27°..... 126

Figure 5.9 Four views of the positive and negative sequence components of the addition of the first and second harmonics (Ap1= 0.35; A_{n1}= 0.15; A_{p2} = 0.15; and A_{n2}=0.05 being equivalent to A1=0.5; B1=0.2; A2=0.2 and B2= 0.1)..... 127

Figure 5.10 Four views of the first harmonic positive and negative sequence components A_{p1}= 0.35 and A_{n1}= 0.15 being equivalent to A1 at 0.5 and B1 at 0.3 127

Figure 5.11 Four views of the second harmonic positive and negative sequence components Ap2 at 0.2 and A_{n2} at 0.05 being equivalent to A2 at 0.2 and B2 at 0.1. 128

Figure 5.12 Four views resulting from the combination of two ellipses having parameters of A1 at 0.5 and B1 at 0.2 and A2 at 0.1 and B2 at 0.2. 129

Figure 5.13 Four views of the positive and negative sequence components with A_{p1}= 0.35 and A_{n1}= 0.15 and A_{p2} = 0.15 and A_{n2} at -0.05 being equivalent to A1 at 0.5 and B1 0.2 and A2 at 0.1 and B2 at 0.2. 129

Figure 5.14 Four views of the second harmonic positive and negative sequence components A_{p2} at 0.15 and A_{n2} at -0.05 being equivalent to A2 at 0.1 and B2 at 0.2. 130

Figure 5.15 Different azimuth views and similar elevation views of two ellipses that produce a cross-over in the contour (Azimuth -View 1 =-90°, View 2 =-69°, View 3 =-48°, View 4 =-10°; Elevation (typical) = -2°)..... 130

Figure 5.16 Diagram showing the 'addition' of two elliptical phasors rotating at ω_1 and ω_2 radians/sec and at orientations of θ_1 and θ_2 132

Figure 5.17 2D and 2DT profile of a 'Triangular' gesture trajectory 133

Figure 5.18 Formation of a triangular trajectory..... 134

Figure 5.19 First three harmonics of 'triangular' gesture trajectory showing 2DT (left) and 2D (right) views all at an orientation equal to the spatial orientation angle (1^{st} harmonic=red, 2^{nd} harmonic=green, 3^{rd} harmonic = blue) 135

Figure 5.20 2DT and 2D views of the harmonics of a shallow concave trajectory showing 1st (red) and 3rd (blue) harmonics at the same orientation, but the 2nd (green) harmonic at a significantly different orientation 136
Figure 5.21 Four views of the synthesis of the 'figure of eight' trajectory from the coefficients of the first three harmonics given in Table 5.7 138
Figure 5.22 A 2DT comparison of original trajectory black, transposed and scaled, with the synthesised trajectory (green). 138
Figure 5.23 2DT views of an arc type gesture trajectory (black on right image) that has an arc or non-planar trajectory characteristic. The third harmonic demonstrates an 'elliptic corkscrew' (blue) as shown on the left-hand picture. The first (red) and second (green) harmonics are shown in the right image. ... 139
Figure 5.24 Original data (cyan circles) and data produced by the IFFT from the first 6 harmonics (black crosses) 140
Figure 5.25 Positive and Negative sequences Magnitude and Phase for 12 harmonics 141
Figure 5.26 Positions of the first 4 harmonics on an Argand diagram for positive (red cross) and negative (black circle) sequence 142
Figure 5.27 The Stopping Tolerance set to within 5 pixels of the Start Coordinates 143
Figure 5.28 Argand Diagram representation of magnitude and phase shift for the first 6 harmonics, when the Stop Tolerance is 2, 5, 10, 15 (first to sixth harmonic, red, green, blue, cyan, magenta and yellow respectively). 144
Figure 5.29 Argand Diagram representations of Magnitude and Orientation Angle (first to sixth harmonic, red, green, blue, cyan, magenta and yellow, respectively). 144
Figure 5.30 Gesture coordinates based on SCM object data 145
Figure 5.31 Gesture coordinates based on manually/visually recorded data 146
Figure 5.32 Gesture coordinates based on SCMI data 146
Figure 5.33 IFFT of SCMI trajectory data using 6 harmonics (black, '+') and original data (cyan, 'o') 147
Figure 5.34 2DT view of the trajectory of five pointing gestures 148

Figure 5.35 Orientation angle 'o', raw phase '.' , and average '*', for six gesturers performing the same gesture (red=1st, green=2nd, blue=3rd, cyan=fourth harmonic) 149

Figure 5.36 The first harmonic orientation angle for the five pointing gestures from one gesturer. 150

Figure 6.1 Schematic diagram of the RBF neural network, with the input layer to the left, RBF hidden layer in the middle and linearly combined with weights to give output f(x) to the right. (Source: Orr, 1996) 154

Figure 6.2 First four harmonics (red, green blue and cyan) of five gestures from six gesturers. 162

Figure 6.3 First four harmonic (red, green blue and cyan) of five gestures from six gesturers – normalised to magnitude of two for first harmonic 163

Figure 6.4 Dendrogram of 30 gestures using Euclidean distance metric and 'ward' linkage method for the first harmonic vector. 165

Figure 7.1 The row and column coordinates for the ten repeated hand raising gestures. 173

Figure 7.2 The first four harmonic orientation angles for ten repeated hand-raising gestures. Harmonics 1-4 are red, green, blue and cyan, shown as 'o' respectively, and average value shown as '*'. Vector magnitude normalised by A_p 174

Figure 7.3 The first four harmonic orientation angles for ten repeated hand-raising gestures. Harmonics 1-4 are red, green, blue and cyan, shown as 'o' respectively, and average value shown as '*'. Vector magnitude normalised to 2. 174

Figure 7.4 A JPEG image showing the right hand (blue '+', 3rd SCM object) about to disappear at the end of the sequence. The left hand (red '+', 1st SCM object) rising for 10 frames. The head is also detected moving (green cross, 2nd SCM object) 175

Figure 7.5 An image from a 'Take Mug' sequence 176

Figure 7.6 Ensemble of 'Take Mug' trajectories of pixel value verses gesture length for row and column coordinate data. 177

Figure 7.7 First six harmonics vectors (red, green, blue, cyan, magenta and yellow respectively) of twenty-one gesturers performing the 'Take-Mug' gesture, normalised to Ap equal to 1 for the 1st harmonic. 179

Figure 7.8 First three harmonics vectors (red, green and blue respectively) of twenty-one gesturers performing the 'Take-Mug' gesture. Normalised to Ap plus An of 1st harmonic equal to 2 179

Figure 7.9 Two clusters of the first harmonic using the Euclidean metric and ward linkage method 182

Figure 7.10 Two clusters of the first harmonic using the 'Euclid' distance metric and the 'single' linkage method that isolate the outlier vector. 182

Figure 7.11 Dendrogram of the second harmonic using the Euclid metric and single linkage 184

Figure 7.12 Graph showing the classification of vectors (red/circle, blue/diamond, green/square and cyan/star) using the Euclid metric and single linkage method for the second harmonic 185

Figure7.13 Dendrogram of clusters for the second harmonic using the 'euclid' metric and 'ward' linkage method. 185

Figure7.14 Graph showing the classification of vectors (red/circle, blue/diamond, green/square and cyan/star) using the Euclid metric and ward linkage method for the second harmonic 186

Figure 7.15 Graph showing the classification of vectors (red/circle, blue/diamond, green/square and cyan/star) for the Euclid distance metric and ward linkage for the third harmonic and showing average values of each cluster (black/star). .. 186

Figure 7.16 A challenging environment for skin-coloured segmentation with inappropriate SCM objects recorded by the green and blue crosses. 191

Figure 7.17 The OSA output, '+', chosen from the two most significant SCM objects (red and green respectively) 192

Figure 7.18 Segmented gesture ready for frequency analysis, showing right hand row and column coordinates with initial search region (red dots), start coordinates (cyan dots) and stop condition (magenta dots). 192

Figure 7.19 Image from the 'whisk' sequence showing gesturer and environmental conditions 194

Figure 7.20 The trajectory coordinates due to the 'whisk' action 194

Figure 7.21 'Whisk' gesture coordinates prepared for frequency analysis..... 195
Figure 7.22 Image from the 'saw-action' sequence showing gesturer and environmental conditions..... 196
Figure 7.23 'Saw-action' gesture coordinates prepared for frequency analysis..... 196
Figure 7.24 2D and 2DT views of the frequency components of the 'Saw-action'. The 'cyan' 4th harmonic component is shown as much greater than the other components..... 197
Figure 7.25 A gestures coordinates (cyan, 'o') with IFFT reconstruction using 6 (black '+') and 4 (red '+') harmonics..... 198
Figure 7.26 2D and 2DT views of the first four frequency components of a gesture with low amplitude oscillation..... 199
Figure A1.1 Measurement of Body Positions for Avatar Design..... 223
Figure A1.2 Skeletal (frame 1) and Outline Figure (frame 295) of avatar Tessa..... 223
Figure A1.3 Tessa the avatar (frame 660)..... 224
Figure A1.4 A frame from the 'Take Mug' sequence..... 224
Figure A1.5 Frames 1 to 9 from the avatar 'Take Mug' sequence..... 226
Figure A1.6 Frames 10 to 18 from the avatar 'Take Mug' sequence..... 226
Figure A1.7 Frames 19 to 27 from the avatar 'Take Mug' sequence..... 227
Figure A1.8 Frames 28 to 36 from the avatar 'Take Mug' sequence..... 227
Figure A2.1 Plots of the reflectance spectra of the back of the hand of various subjects (Source: Angelopoulos, 2001)..... 229
Figure A2.2 The reflectance spectrum of human skin compared with the absorption spectrum of oxygenated haemoglobin (Source: Angelopoulos, 2001)..... 230
Figure A2.3 The same scene illuminated by four different illuminants (Tungsten, White Fluorescent, D50 and D65)..... 232
Figure A2.4 Position of the 7x7 sample on the doll's arm shown in red..... 232
Figure A2.5 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in the 'Take Mug' sequence for a plus/minus two standard deviation from the mean, based on the three samples..... 234
Figure A2.6 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the 'Take Mug' sequence..... 235

Figure A2.7 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in a complex scene with even illumination for a plus/minus two standard deviation from the mean, based on the three samples. 236

Figure A2.8 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the Complex Scene with Even Illumination. 237

Figure A2.9 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in a Scene with Low Illumination and Poor White Balance for a plus/minus two standard deviation from the mean based on the three samples. 238

Figure A2.10 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the Scene with Low Illumination and Poor White Balance. 239

Figure A2.11 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in a Challenging Environment and Poor White Balance for a plus/minus two standard deviation from the mean based on the three samples. 240

Figure A2.12 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the Scene with a Challenging Environment and Poor White Balance. 241

Figure A2.13 Image '\data\ScenarioA1\Cam1\image16511.jpg' from the PETS database..... 242

Figure A2.14 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in the PETS sequence for a plus/minus two standard deviation from the mean based on the three samples. 242

Figure A2.15 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the PETS sequence..... 243

Figure A2.16 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the PETS sequence to include all skin-coloured regions. 244

Figure A3.1 Images 3 to 11 of the Low Illumination and Poor White Balance sequence showing the position of the first three SCM objects (red=1st, green=2nd, blue=3rd)..... 246

Figure A3.2 Images 12 to 19 of the Low Illumination and Poor White Balance Sequence showing the position of the first three SCM objects (red=1st, green=2nd, blue=3rd) 246

Figure A3.3 Frame 250 showing the centre of gravity of the three SCM objects (red, green and blue crosses) on the head using just the Hue for skin-colour segmentation. 247

Figure A3.4 Frame 250 showing the centre of gravity of the three SCM objects (red, green and blue crosses) on the head using Hue-Saturation for skin-colour segmentation. 248

Figure A3.5 2D and 3D views of the first three SCM objects (red, green and blue, respectively), for experiment E1 248

Figure A3.6 2D and 3D views of the first three SCMI objects (red, green and blue respectively), for experiment E1 249

Figure A3.7 2D and 3D views of the first three SCME objects (red, green and blue, respectively), for experiment E1 249

Figure A3.8 2D and 3D views of the first three SCMEI objects from Hue mask (red, green and blue, respectively), for experiment E1. 250

Figure A3.9 Experimental conditions E1 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position) 250

Figure A3.10 Experimental conditions E4 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position) 251

Figure A3.11 Experimental conditions E5 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position) 251

Figure A3.12 Experimental conditions E8 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand 252

Figure A3.13 Experimental conditions E1 for SCMI data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position) 253

Figure A3.14 Experimental conditions E5 showing 1st (red circle) and 2nd (green diamond) SCM object coordinates. The tracking output is signifies the right hand by the black '+' and the left hand by the blue 'x'. The cyan and magenta dots represent the visually obtained left and right hand coordinate positions, respectively. 254

Figure A3.15 The positioning of the first three most significant SCM objects (red, green and blue, respectively) on frames 16960 to 16968. 255

Figure A5.1 (a) A 2-D closed contour. (b, c) Periodic functions $X(l)$ and $Y(l)$ for the contour of (a) (Source: Lin et al.) 263

Figure A5.2 Different starting points due to different orientations (Source: Lin et al.) 264

Figure A5.3 The rotation and starting phase of an ellipse. Source (Lin et al.) 265

Figure A5.4 Three harmonic elliptic descriptions. Source (Lin et al.) 265

Figure A5.5 Pictures of 4 views of the summation of two ellipses of different frequencies with an offset. The top left picture shows the spatial domain representation; the top-right shows the change of 'y' with time, clearly showing the offset bias; the bottom-left picture shows the change of 'x' with time and the bottom right shows a 2DT view..... 266

Figure A5.6 Different azimuth but the same elevation of two ellipses with offset (Azimuth -View 1 =30°, View 2 =20°, View 3 =10°, View 4 =0°; Elevation = 30°) 267

Figure A5.7 Four views (top-left spatial domain, x-y; top-right, 'y' vs. time; bottom-left, 'x' vs. time; bottom-right, 2DT domain) of the first harmonic. The starting point-and end-point of the time sequence indicated as 'o' and '*' (blue) respectively. 268

Figure A5.8 Four views (top-left spatial domain, x-y; top-right, 'y' vs. time; bottom-left, 'x' vs. time; bottom-right, 2DT domain) of the third harmonic. Rotating positive sequence (red) and negative (black) sequences and the resulting (blue) ellipse. 268

Figure A5.9 Four views of combining the first, third and fifth harmonics. Rotating positive (red) and negative (black) sequences and the resulting (blue) ellipse. The starting point-and end-point of the time sequence indicated on all pictures as 'o' and '*' (blue) respectively. 269

Figure A5.10 2DT and 2D views of the harmonics of a shallow arc or concave Trajectory 271
Figure A5.11 2DT and 2D views of the harmonics of a deep arc or concave trajectory. 272
Figure A5.12 2DT and 2D views of the harmonics of a convex trajectory 273
Figure A5.13 2DT and 2D views of an elliptical trajectory 274
Figure A5.14 2DT and 2D views of a 'figure of eight' trajectory 275
Figure A5.15 2DT and 2D views of the harmonics of an oscillatory trajectory 276
Figure A5.16 2DT and 2D Views of the harmonics of a real gesture trajectory 277
Figure A5.17 A third harmonic, single 'elliptic-corkscrew' 278
Figure A5.18 A fourth harmonic, single 'elliptic-corkscrew' 278
Figure A5.19 Row and column coordinates of a trajectory generated using SCM objects. 281
Figure A5.20 Row and column coordinates of a trajectory generated using SCMI objects. 282
Figure A5.21 Row and column coordinates of a trajectory recorded visually 283
Figure A6.1 Diagram showing object number and coordinates for the five objects used in the clustering examples to show the differences between distance metrics and linkage methods. 302
Figure A6.2 Distances between objects for 'single' linking and Euclidean distance metric 303
Figure A6.3 Dendrogram for 'single' linking and Euclidean distance metric 304
Figure A6.4 Distances between objects for 'complete' linking and Euclidean distance metric 305
Figure A6.5 Dendrogram for 'complete' linking and Euclidean distance metric 306
Figure A6.6 Distances between objects for 'average' linking and Euclidean distance metric 307
Figure A6.7 Distances between objects for 'average' linking and Euclidean distance metric 308
Figure A6.8 Distances between objects for 'centroid' linking and Euclidean distance metric 309
Figure A6.9 Dendrogram for 'centroid' linking and Euclidean distance metric 310

Figure A6.10 Distances between objects for 'ward' linking and Euclidean distance metric 311
Figure A6.11 Dendrogram for 'ward' linking and Euclidean distance metric 312
Figure A6.12 Dendrogram using 'single' linking with Mahalanobis distance metric 313
Figure A6.13 Dendrogram using 'complete' linking with Mahalanobis distance metric 314
Figure A6.14 Dendrogram for 'average' linking with Mahalanobis distance metric 315
Figure A6.15 'City Block' distance metric, 'single' linkage method 316
Figure A6.16 'City Block' distance metric, 'complete' linkage method 316
Figure A6.17 'City Block' distance metric, 'average' linkage method 317
Figure A6.18 'City Block' distance metric, 'centroid' linkage method 317
Figure A6.19 'City Block' distance metric, 'ward' linkage method 318
Figure A6.20 'Euclidean' distance metric, 'single' linkage method 318
Figure A6.21 'Euclidean' distance metric, 'complete' linkage method 319
Figure A6.22 'Euclidean' distance metric, 'average' linkage method 319
Figure A6.23 'Euclidean' distance metric, 'centroid' linkage method 320
Figure A6.24 'Euclidean' distance metric, 'ward' linkage method 320
Figure A6.25 'Mahalanobis' distance metric, 'single' linkage method 321
Figure A6.26 'Mahalanobis' distance metric, 'complete' linkage method 321
Figure A6.27 'Mahalanobis' distance metric, 'average' linkage method 322
Figure A7.1 Illustration of a gesturer stationary (frame 2) and in the process of enacting the five gestures (frames 32, 62, 92, 122 and 152) 323
Figure A7.2 The row and column coordinates of the continuous trajectory of five pointing gestures for one gesturer 324
Figure A7.3 Frames 15 to 23 of gesture type 1 324
Figure A7.4 Frames 24 to 32 of gesture type 1 325
Figure A7.5 Frames 33 to 41 of gesture type 1 325
Figure A7.6 2D and 3D representation of the first three harmonic components of the fifth PETS hand raising gesture. 327
Figure A7.7 Distribution of the first three harmonic vectors for the ten hand raising gestures. 330

Figure A7.8 Distribution of the vector magnitudes for the ten hand raising gestures. 330
Figure A7.9 Alternate frames 2882 to 2898 from a 'Take-Mug' gesturer..... 331
Figure A7.10 Alternate frames 2900 to 2916 from a 'Take-Mug' gesturer..... 331
Figure A7.11 Alternate frames 2918 to 2934 from a 'Take-Mug' gesturer..... 332
Figure A7.12 Alternate frames 2936 to 2952 from a 'Take-Mug' gesturer..... 332
Figure A7.13 Alternate frames 2954 to 2970 from a 'Take-Mug' gesturer..... 333
Figure A7.14 Alternate frames 2972 to 2988 from a 'Take-Mug' gesturer..... 333
Figure A7.15 Distribution of the first six harmonics of the twenty-one 'Take Mug' gestures. Normalised by A_p =1 335
Figure A7.16 Visual grouping of gestures similar to gesture A. 336
Figure A7.17 Visual grouping of gestures similar to gesture G. 337
Figure A7.18 Visual grouping of gestures similar to gesture M..... 338
Figure A7.19 Visual grouping of gestures similar to gesture K. 339
Figure A7.20 Alternate frames 451 to 457 of a 'whisk' gesturer 340
Figure A7.21 Alternate frames 469 to 485 of a 'whisk' gesturer 340
Figure A7.22 Alternate frames 451 to 457 of a 'whisk' gesturer 341
Figure A7.23 Alternate frames 451 to 457 of a 'whisk' gesturer 341
Figure A7.24 Alternate frames 451 to 457 of a 'whisk' gesturer 342
Figure A7.25 Alternate frames 936 to 952 of a 'saw-action' gesturer..... 343
Figure A7.26 Alternate frames 954 to 970 of a 'saw-action' gesturer..... 343
Figure A7.27 Alternate frames 972 to 988 of a 'saw-action' gesturer..... 344
Figure A7.28 Alternate frames 990 to 1006 of a 'saw-action' gesturer..... 344
Figure A7.29 Alternate frames 1008 to 1024 of a 'saw-action' gesturer..... 345
Figure A7.30 Alternate frames 1026 to 1042 of a 'saw-action' gesturer..... 345
Figure A7.31 Alternate frames 1044 to 1060 of a 'saw-action' gesturer..... 346

List of Tables

Table 3.1 Rank ordered size filtered motion data 77
Table 3.2 Rank ordered size filtered gesture objects 78
Table 5.1 Proportion of different movement primitives (Source, Gibet 2001)..... 115
Table 5.2 Comparison of ellipse definition coefficient A and B with A_p and A_n 126
Table 5.3 Triangular trajectory magnitude and phase information for the first 7.... 133
harmonics, based on 31 points 133
Table 5.4 Triangular trajectory magnitude and phase information based on 31 points
..... 133
Table 5.5 Triangular trajectory magnitude and phase information based on original
49 points..... 135
Table 5.6 Harmonic values of a shallow concave trajectory 137
Table 5.7 The first three harmonic coefficients used to synthesis the original
waveform 137
Table 5.8 Table of complex data representing the first 4 harmonics of a gesture
trajectory 140
Table 5.9 Table of magnitude and phase representing the first 4 harmonics of a
gesture trajectory 141
Table 5.10 Table of magnitude and orientation angle separated from phase for the
first 4 harmonics of a gesture trajectory 142
Table 5.11 Gesture Number and Action 148
Table 5.12 Orientation Angle, θ and Positive and Negative Sequence Magnitude and
Phase, ϕ representation for gesture A1 148
Table 5.13 Comparison of all gesturers orientation angle for the first two harmonics
of gesture 1 149
Table 6.1 Target gestures made from the average of all 6 gestures, misclassification
shown bold, near miss shown after '/'. Harmonic pair = 1..... 161
Table 6.2 Target gestures made from the average of all 6 gestures, misclassification
shown bold, near miss shown after '/'. Harmonic pair = 12..... 161
Table 6.3 A_p, A_n and Orientation angle for the first 4 harmonics normalised by A_n,
for the gesture A1. 163

Table 6.4 Values of Table 6.3 normalised to the first harmonic magnitude of 2 showing real and imaginary coordinates. 163

Table 6.5 Cophenetic correlation coefficient values when comparing of distance metrics and linkage methods for the first harmonic vector, for normalisation arbitrarily chosen as A_p=1 164

Table 6.6 Cophenetic correlation coefficient values when comparing of distance metrics and linkage methods for the first harmonic vector, for normalisation when A_p or A_n is the greatest 164

Table 6.7 Cophenetic correlation coefficient values when comparing of distance metrics and linkage methods for the first harmonic vector, for normalisation of the first harmonic to value 2. 164

Table 6.8 Classification of gestures for each gesturer using 'euclid' distance metric and 'ward' linkage method (bold shows miss-classification)..... 166

Table 6.9 Classification of gestures for each gesturer using 'euclid' distance metric and 'single' linkage method (bold shows miss-classification). 166

Table 6.10 Gesture Number and Action Alignment 167

Table 6.11 Comparison of PNN method (P) and Clustering method (C) for classification using one harmonic, (bold shows miss-classification). 167

Table 7.1 Frequency Response of Sequence A of 'Take Mug' suite of Experiments, original length of 63..... 177

Table 7.2 Frequency Response of Sequence A of 'Take Mug' suite of Experiments, reduced to 32 from original length of 63. 178

Table 7.3 Classification of the gestures into the four sub-classes, α, β, γ and δ of the 'Take Mug' gesture using a PNN and least squares calculations ranks the result to the nearness of the other targets. The first three orientation angles are compared to the classification. 181

Table 7.4 Comparison of distance metrics and linkage methods by the cophenetic correlation coefficient with the revised normalisation method for the first harmonic. The largest coefficient is shown in bold. 183

Table 7.5 Comparison of distance metrics and linkage methods by the cophenetic correlation coefficient with the revised normalisation method for the second harmonic. The largest coefficient is shown in bold. 183

Table 7.6 Comparison of distance metrics and linkage methods by the cophenetic correlation coefficient with the revised normalisation method for the third harmonic. The largest coefficient is shown in bold. 184
Table 7.7 Target Classes generated for the second and third harmonics clusters using the 'euclid' distance metric and 'ward' linkage method, resulting in Five Target Classes (US=Unspecified as is a single entity) 187
Table 7.8 Target and Target Classes and associated cluster of the Second and Third harmonic using the Euclidean distance metric. 188
Table 7.9 Target coordinate data – Euclidean distance metric (Italic data not used) 188
Table 7.10 Classification of twenty-one 'Take Mug' gestures using the PNN/ clustering technique) using Euclidean distance metric and 'ward' linkage method and compared to the visual classification (bold shows the difference in the metric methods and / show the next nearest target). 189
Table 7.11 Comparison of classification techniques for 21 'Take Mug' gestures (bold shows difference between the two classifications techniques) 190
Table 7.12 'Whisk' frequency content, showing the prominence of the seventh harmonic (bold). 195
Table 7.13 'Saw-action' frequency content, showing the prominence of the fourth harmonic (bold) 197
Table 7.14 Frequency content of a low amplitude oscillation gesture 199
Table A2.1 RGB and HSV values in Tungsten Light 233
Table A2.2 RGB and HSV values in White Fluorescent Light 233
Table A2.3 RGB and HSV values in Graphics Art D50 Light 233
Table A2.4 RGB and HSV values in Textile Dye D65 Light 233
Table A3.1 Experimental Combinations of H (Hue), HS (Hue-Saturation), Hole 'Fill' and 'Opening' 245
Table A3.2 Comparison of the performance of experiments 1, 4, 5 and 8 for positioning of the 1st 'SCM' object 252
Table A3.3 Allocation of the first three SCM objects to subjects 6, 5 and 4 (left to right) and place on body (rh = right hand, lh = left hand and f = face) of the gesturer 6. 259
Table A4.1 Calculation of ratios for normalisation of gesture length to 64. 262

Table A5.1 Calculation of harmonic components for data 1, 0, 0, 1 263
Table A5.2 Harmonic Values of a Shallow Concave Trajectory 271
Table A5.3 Harmonic values of a deep arc or concave trajectory 272
Table A5.4 Harmonic values of a deep convex trajectory 273
Table A5.5 Harmonic values of an elliptic trajectory 274
Table A5.6 Harmonic values of 'figure of eight' trajectory 275
Table A5.7 Harmonic values of an oscillatory trajectory 276
Table A5.8 Harmonic values of a real gesture trajectory 277
Table A5.9 Harmonics with stopping Tolerance 2; Gesture Length 37 279
Table A5.10 Harmonics with stopping Tolerance 5; Gesture Length 36 279
Table A5.11 Harmonics with stopping Tolerance 10; Gesture Length 34 280
Table A5.12 Harmonics with stopping Tolerance 15; Gesture Length 32 280
Table A5.13 Harmonics generated using SCM objects. 281
Table A5.14 Harmonics generated using SCMI objects. 282
Table A5.15 Harmonics generated from visually recorded coordinates 283
Table A6.1 Object number and coordinate value 301
Table A6.2 Euclidean distance matrix derived from the five objects 301
Table A6.3 Distances of objects using 'single' linking and Euclidean distance metric 303
Table A6.4 Distances of objects using 'complete' linking and Euclidean distance metric 305
Table A6.5 Distances of objects using 'average' linking 307
Table A6.6 Distances of objects using 'centroid' linking and Euclidean distance metric 309
Table A6.7 Distances of objects using 'centroid' linking and Euclidean distance metric 311
Table A6.8 Mahalanobis distance matrix derived from the five objects 313
Table A6.9 Distances of objects using 'single' linking with Mahalanobis distance metric 313
Table A6.10 Distances of objects using 'complete' linking and with Mahalanobis distance metric 314
Table A6.11 Distances of objects using 'average' linking with Mahalanobis distance metric 315

Table A7.2 Orientation angles (θ) for 6 harmonics and 10 gestures, data obtained visually. 328
Table A7.3 Magnitude, (M = positive + negative sequence value) for 6 harmonics for 10 gestures, data obtained visually. Normalised by A_p =1. 328
Table A7.4 Orientation angles (θ) for 6 harmonics for 10 gestures, data obtained automatically. 329
Table A7.5 Magnitude (M = positive + negative sequence value) for 6 harmonics for 10 gestures, data obtained automatically. Normalised by A_p =1 329
Table A7.6 Description of the characteristics and differences of each individual 'Take Mug' gesturer. 334
Table A7.7 Average and Standard Deviation (std.) of twelve harmonics from the twenty one 'Take Mug' gestures. 335
Table A7.8 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture A. Normalised by A_p =1 336
Table A7.9 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture G. Normalised by A_p =1 337
Table A7.10 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture M. Normalised by A_p =1 338
Table A7.11 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture K. Normalised by A_p =1 339
Table A7.12 A gesturer's response to gesture stimuli 346

Acknowledgements

The work presented in this thesis has been carried out whilst working at Buckinghamshire Chilterns University College and I would like to thank the authorities for their support to enable me to do this work.

My special thanks go to my supervisor Tim Ellis for his support, guidance and critical comments, suggestions and discussions which have made me question and enhance my work.

I am indebted to friends, colleagues and family for tolerating my wonderment about gestures and their origin. I am especially grateful to Margaret, my wife's , support in this endeavour.

Author's declaration of previously published work

Harding, P. R. G. and Ellis, T. J. (2004), Recognizing Hand Gesture using Fourier Descriptors, *International Conference Pattern Recognition,* **3**, 286-289

Harding P R G and Ellis T J (2003), An Improved Skin Colour Segmentation Technique for Hand Gesture Tracking, Poster *at International Gesture Workshop*, Genoa, Italy.

Harding, P. (1999), Investigations into Skin Colour as a Feature for Detecting Hand Gestures, *Technology Letters*, **3**, 1, 1999, 34-43.

Abbreviations

Terminology and abbreviations used within this thesis are listed here in alphabetical order and not the order in which they appear.

- **Apraxia** the deficit in the ability to understand an action or to perform an action in response to verbal command or in imitation.
- **Avatar** a synthesised computer generated character displaying human characteristic.
- **ANN** Artificial Neural Network
- **ASL** American Sign Language
- **Back-propagation** a method of supervised learning where the errors in the output nodes are used to change the weights of the network.
- **Babbling** noises that sound like words by a baby before speech evolves.
- **Beats** an oscillatory gesture characteristic.
- **Bi-phasic** a gesture definition related to the 'beat' type of gesturing characteristic.
- **CIE** Committee Internationale de l'Eclairage.
- **CMYK** Cyan Magenta Yellow and Black.
- **Deictic Gesture** A 'pointing' gesture.
- **D50** Graphics Art Lighting Standard.
- **D65 Textile** Dye Lighting Standard.
- **EM** Expectation Maximisation.
- **Feed-Forward** input vector applied to input layer of network and fed through the network to the output layer.
- **FFT** Fast Fourier Transform
- **FSL** French Sign Language
- **HamNoSys** Hamburg Notation System - a notation for transcribing signing gestures
- **HMM** Hidden Markov Model.
- **HSI** Hue Saturation Intensity.
- **HSV** Hue Saturation Value.
- **Humanoid** an animated model with human characteristics

- **Iconic Gesture** a gesture typically representing for example a book or film.
- **IFFT** Inverse Fast Fourier Transform
- **Makaton a** signing language for disabled children.
- **Manikin** a model with human appearance.
- **Markov Model** a probabilistic pattern-matching technique that models a time sequence as the output of a stochastic or random process.
- **Metaphoric Gesture** an abstract gesture.
- **Metacarpus a** bone at the back of the hand.
- **MAD** Mean Absolute Difference.
- **MDVI** Multiply Disabled and Visually Impaired.
- **MHI** Motion History Images.
- **MLP** Multi-Layer Perceptrons.
- **MLFN** Multi-Layer Feedforward Network.
- **MLP NN** Multi Layer Perceptron Neural Network.
- **MSD** Mean Squared Difference.
- **NN** Neural Network of processing units linked together by weighted connections.
- **OSA** Object Selection Algorithm.
- **PCA** Principal Component Analysis.
- **PETS** Performance Evaluation of Tracking and Surveillance systems.
- **PNN** Probabilistic Neural Network.
- **RBF** Radial Basis Function.
- **RGB** Red Green Blue.
- **RNIB** Royal National Institute for the Blind.
- **RNN** Recurrent Neural Networks.
- **SCM object** A gesture object produced by the logical AND of a Skin-Colour Mask and a Motion Mask.
- **SCME object** An object produced by the logical AND of a Skin-Colour Mask, a Motion Mask and an Edge Template Mask.
- **SCMEI object** An object indicated by the skin-colour mask being congruent with the SCME object.

- **SCMI object** An object indicated by the skin-colour mask being congruent with the SCM object.
- **SHI** Skin History Image.
- **SWOP** Standard Web Offset Press.
- **TDNN** Time-Delay Neural Network.
- **TDRBF** Time-Delay Radial Basis Function.
- **TESSA,** the Text and Sign Support Assistant
- **TLFN** Focused Time Lagged Feed Forward Network.
- **Tri-phasic** a definition of the three divisions (preparation, stroke and retraction) of a gesture.
- **ViSiCAST Virtual Signing: Capture, Animation, Storage and Transmission,** a project to facilitate access, by deaf citizens, information and services in sign language using computer generated humans or avatars.
- **VLMM** Variable Length Markov Model
- **1D** One Dimensional
- **2D** Two Dimensional
- **2DT** Two Dimensional and Time
- **3D** Three Dimensional

1. Introduction

The importance of human gesture has become more readily understood since the publication of books such as 'The Naked Ape', 'Manwatching' and 'Gestures' (Morris et al., 1967, 1977, 1979). Morris et al., says that the science of human gesture has been greatly underestimated compared to the number of people involved in linguistics or in the analysis of language. It has been considered that gestures are a trivial, insignificant form of human communication, but it is becoming clear that social intercourse depends heavily on the actions, postures, movements and expressions of the talking body. Morris et al. (1979) claims that gestural information is more important than words when it comes to indicating changing mood or emotional state. The analysis of gesturing has lagged behind the science of linguistics. This maybe because gesturing is difficult to record and to explain in words. The value of gesturing as an important fact of non verbal communications has been given inadequate formal recognition or annotation.

People effortlessly read other people's gestures and emotional states but only a few have had the ability to record the information as well. Thomas Hardy (1891) recognised the changes to facial colour that he suggests indicate the character's mood.

"Tess cried, and the colour upon her cheeks spread over her face and neck. In a moment her eyes grew moist, and her glance dropped to the ground."

To detect the change in the colour of Tess's cheek would have been difficult to do by machine a few years ago. Sanguine (2000) states that one of the factors that have inhibited the development of colour imaging in the past has been limited computer memory and processing speed. Another factor has been the high cost of cameras and displays. These factors meant that although digital image processing has a history of about forty years, it was difficult to experiment with algorithms or have the motivation to test new ideas. Today a domestic or Notebook personal computer has sufficient space to store large number of images with a sufficient processing speed to process images in reasonable speed and even real time.

Jane Austen (1813) described the use of bodily movements to convey meaning: -

"Elizabeth shook her head over this letter."

In this example Austen conveys the metaphorical intention as well the literal sense of head shaking. Modern authors like Ondaatje (1992), tend to use gesture examples more freely: -

"He pats his chest as though looking for his pass, ….."

Kelly et al. (1999) comment that most theories of pragmatics take as the basic unit of communications the verbal or written utterances, but have overlooked the fact that important information about an utterance's meaning can be conveyed by non verbal communications.

Over the last twenty years, the scientific community has been interested in the study of gestures using the computer. In this country, the British Machine Vision Association was formed in 1990. It holds annual conferences which often attract papers concerned with human tracking and face recognition. Although there are many other international organisations that recognise the work of the gestural community, a significant organisation is the International Gesture Workshop that is held every two years. The International Gesture Workshops are interdisciplinary events for those researching gesture-based communications for those wishing to meet and exchange ideas across disciplines. A focus of these events is a shared interest in using gesture and sign language in human-computer interaction.

For example, Daugman (1997) observed that people interact with computers as well as people. The ability to develop machine interfaces that have characteristics similar to the human skills of recognizing facial features, gestures and speech is preferred, rather than humans develop and learn machine skills.

This thesis explores the nature and application of the particular human predisposition of gesturing. The thesis reviews approaches that have been made to interpret and mimic gestures using the power of the computer to then describe a system for analysis and recognising hand-gestures for a range of gestures and gesture stimuli.

1.1. The origin of gesture

How does gesturing in humans begin?

Natural gesturing is as evident as babbling in normal babies at two months. This can be seen clearly in deaf babies. Deaf babies are seen to babble with their hands at the same time (Siple, 1978). The appearance of one-word utterances and then two-word strings of hearing children, occur at the same time as do the appearance of single sign and two-sign strings in deaf children who are in a signing environment.

Further evidence that the ability to communicate is inborn has come from watching the gestures of deaf children. (Goldin-Meado and Mylander, 1998*)*. Four American and four Chinese children aged between three and five months were observed. Sign language was unknown to the children, but they managed to convey quite complex messages by action and gesture. It was found that the syntax of the children's gesture consisted of sentences rather than just words. The syntax for both the American and Chinese children was similar. It was concluded that the sentence construction must be 'hardwired' at birth rather than learnt. Interestingly, this research showed that the gestures of the children were different to their mother's gestures, indicating that they had invented their own repertoire of gestures.

Furthermore, it is argued that language evolved from manual gestures, gradually incorporating vocal elements (Corballis, 2003). It is also suggested that the emergence of language was from manual to facial gesture and then incorporated sound. Although modern sign language does not necessarily resemble gesture language used by our ancestors it is observed that with American Sign Language facial gesture generally conveys context, whereas manual gestures supply content. Corballis observes that the strong predominance of right-handedness appears to be a uniquely human characteristic, although the left-cerebral dominance for vocalisation

appears in many species, including frogs, birds and mammals. Furthermore, right-handedness may have arisen because of an association between manual gestures and vocalization in the evolution of language. Interestingly, impairment in praxis functioning is common after a stroke, most frequently when the left hemisphere is affected (Koski et al., 2002). Apraxia is defined as a deficit in the ability to understand an action or to perform an action in response to verbal command or in imitation e.g. wave goodbye, pantomime use of a hammer.

The importance of gesture accompanying speech has been found to be influential in conveying information (Kelly et al., 1999). It has been found that people are more likely to interpret an utterance as an indirect request when speech is accompanied by a relevant pointing gesture than when speech or gesture was presented alone. The combination of gesture with vocal utterance adds to the communication.

1.2. Human-Computer Interface Applications

A useful introduction to computer vision techniques being applied to gesturing is made by Daugman (1997). This overview explains how face and gesture recognition is an effortless aspect of human behaviour. Although some people have gained effective interactive skills with computers, he suggests that it would be better if machines developed more human-like skills to recognise faces, gestures and speech, rather than humans acquire machine-like skills. The main factors that determine the performance of the face or other recognition system are that the inter-class variance should be large and the intra-class variance should be small.

The example given is for faces. Different faces should generate face codes that are as different as possible from each other, while different images of the same face should generate similar codes. The captures of facial images are dependent on pose, perspective, angle, illumination, age, cosmetics, adornments and expression. The main problem is to determine what characteristics to extract for analysis, recognition and classification.

To this end, several researchers have looked into hand gesture recognition (Pavlovic et al., 1997). They report on tracking methods where there has been a constraint to natural gesture using a uniquely coloured glove, or marks on hands and fingers, or the use of an electronic glove as an interface to the computer. They also report on the many varied approaches that are being tried out for recognising gesture in natural settings.

This research has potential benefit for the deaf and disabled. Helping deaf and disabled people to interpret gestures has obvious uses for improved communications between signing and non-signing people. Work has been carried out in this area, usually using American Sign Language (ASL) (Starner and Pentland, 1995), (Lockton and Fitzgibbon, 2002).

Gesture capture can also be used to control equipment (Howell and Buxton, 1998). For example, instead of holding an infrared remote handset, a camera can be placed on a television set to monitor the audience. Changing channel and volume can take place by the interpretation of gesture. These ideas can be further extended into the automatic control of cameras in a teleconferencing situation. Cameras can be

automatically trained on the people talking or show the gestures applicable to useful speech.

There is a further development as a result of research studies in the area. The University of Michigan Laboratory for Human Motion Simulation (HUMOSIM) is developing mathematical models for human movement prediction. This information is being gained from a variety of workplace and consumer environments. Much of the impetus of this work is to address workplace related musculoskeletal injuries. From the large number of trials so far recorded, research is being carried out into characterizing movement patterns and the limitation of a range of group categories, for example: age, gender, height, body size and so on. (older, younger, male, female, large, small). Howell and Buxton (2003) have used hand trajectory data from the HUMOSIM project, 'Terminal Hand Orientation and Effort Reach Study, 2000', to learn gesture and identity from different individuals.

The field of Human-Computer Interface (HCI) has become an important aspect of research, because of the growing importance and popularity of the Graphical User Interface (GUI) and Virtual Reality (VR) systems. Gibet et al. (2001) observes that the massive development of human-computer interaction has resulted in new systems that try to take advantage of the expressive power of gestures. The latest interfaces are more natural and give rise to a number of virtual reality applications. This improves the ability to capture body movements, recognise and interpret human actions so as to animate virtual humans or avatars.

The animation of deaf signing gestures (Kennaway, 2001) is supported by the ViSiCAST project. The aim of the project (Elliott et al., 2000) is to provide deaf citizens with support in the areas of broadcasting, face-to-face transactions, and the World-Wide Webb (WWW) for information and services in the preferred medium of sign language. A key feature of this work is of the use of computer-generated virtual humans, or avatars. The ViSiCAST project has developed from previous projects (Lincoln et al., 2001). A legible deaf-signing virtual human (Pezeshkpour et al., 1999), used a signing avatar system based on motion capture. Motion capture requires the generation of data files of body, arms, hands and face to be recorded for a lexicon of signs. The merit of this system is its greater capacity for authenticity. The disadvantage of this technique is the substantial work involved in setting up and calibrating equipment to record the large number of signs for a lexicon. It is, therefore, quite a complex task to modify captured motions.

The construction of an avatar requires the synthesis of human characteristics to mimic human characteristics successfully. This work is complementary to the work of gesture analysis and the two fields are moving closer together. The recorded motions of a human, displayed by an avatar of a different character can be easily discerned by people (Kennaway, 2001). This is a form of gait recognition or more generally biometrics, where the aim is to identify people notably in the areas of security and surveillance. Gait recognition is an attractive technique as it is non-invasive and does not require the subjects' cooperation or permission. A pertinent example of a gait recognition technique is described by Mowbray and Nixon (2003). This technique models the full movement and deformation of the body, rather than a specific body part.

Disabled children who cannot speak easily are often taught to communicate by gesture stimuli. This can be well organised by methods such as the Makaton vocabulary (Grove and Walker, 1990), but in some cases it is by a very intensive touching and feeling interaction with a skilled tutor. These gesture stimuli are some very basic communication techniques that both disabled and 'normal' people use and understand. The study of how MDVI (Multi-Disabled, Visually Impaired) children are taught to communicate is explained by the RNIB (2004) and can be very instructive as to the role of gestures in communications. The communications programme first develops movement/interaction through individual movement sessions. The second part develops children's own personal gestures. It finally provides a language model and an adapted sign system. Natural gesture arises from the child using familiar movement patterns, that he is already aware of and which he then learns also has a symbolic aspect that adds to its meaning to him. From simple natural gesture of clapping hands or banging a table, for example, objects can be linked to the gesture to give meaning. For instance a mug would indicate a drink and a sponge a bath. Objects are also used to signify particular people and signify place and time. Children then go on to learn a rudimentary sign language and some basic sounds that are easy to interpret.

Human emotions have been studied, probably since civilisations have existed, but it is only recently that the subject has been linked to Human-Computer interaction. Pickard (2004) and others' current research concerns human emotions as part of the Affective Computing Research Area at MIT (Massachusetts Institute for Technology). The origin of human emotion is not clear. There are conflicting theories, a classical 'chicken and egg' situation. The 'James-Lange' theory (Wozniak, 2004) suggests that actions precede emotions and then the brain interprets actions as emotions. Whereas, the Cannon-Bard (1927) theory is an opposing view stating that emotion is felt first. The actions follow from cognitive processing. The general conclusion is that emotion involves a dynamic state that has both physical and cognitive events. Some recent research was undertaken, constructing 'An Affective Tutor', which is an agent that senses the affective states such as, boredom or anxiety for example. It has the capability to adjust its response to the user, in accord with the user's state. Another area concerns the 'Computer Response to user Frustration' in which a human-computer interaction agent was designed to support users with consideration for their ability to recover from negative emotional states. The Affective Computing group at MIT (2004) are interested in developing technologies to assist in the development of human emotional intelligence: -

"Our approach, grounded in findings from cognitive science, psychology, neuroscience, medicine, psychophysiology, sociology, and ethics, is to develop engineering tools for measuring, modelling, reasoning about, and responding to affect. Thus, we develop new sensors, algorithms, systems, and theories that enable new forms of machine intelligence as well as new forms of human understanding. Many of the challenges we face cannot be solved with existing engineering tools; consequently, we also work at the frontiers of research in machine learning, pattern recognition, signal processing, computer vision, speech analysis, sensor design, human-centred and value-centred design, and more."

The work in this thesis is grounded in signal processing, computer vision and pattern recognition techniques that culminate in identifying some psychological and

emotional states of the gesturer. The gesturing researched is based on gesture stimuli which in evolutionary terms precedes any formalised sign language and is often one-handed. This type of gesture has not been explored in much detail because of the difficulty of explaining or recording the process. But it is extensively used unconsciously by 'hearing' people and used more consciously as a rudimentary sign language by people with hearing difficulties. Although the main action of the gesture is conducted by the dominant hand, the non-dominant hand can act as an indicator of the gesturer's psychological or emotional state. The research shows that gesture stimuli can be very individual to the gesture maker. These intra-class variations could be classed as a form of gait recognition. Kendon (1986) described gestures as 'bi-phasic' or 'tri-phasic'. These definitions are based on the temporal nature of a hand gesture that can be divided up into three parts: the preparation phase in which the hand moves from a resting position; the stroke phase and the retraction phase when it returns to the resting position. A 'bi-phasic' or 'beats' gesture is an oscillatory gesture which is seen in a typical 'finger-wagging' type of gesture.

In this investigation of gesturing the gesture trajectory is modelled as an aperiodic waveform showing movement in two dimensions and time (2DT). The coordinates of the hand are derived from a combination of skin-colour and motion cues. Fourier analysis techniques are applied to the trajectory to produce frequency or harmonic components to characterise the gesture. Much of the stimulus for this work came from Fourier Descriptor techniques (Kuhl and Giardina 1982; Lin and Huang, 1987; Lin and Jungthirapanich, 1990) that are used for object recognition in the spatial domain. The analysis of gesture trajectories was investigated in the time domain and models and experiments were produced that confirmed the equations describing the 2DT (Two spatial dimensions and a time dimension) motion of the hand. Interestingly the Fourier analysis technique showed that in the 'stroke' phase of the gesture some gestures showed a high oscillatory component not normally seen in signing. As a result it is suggested that the definition of a tri-phasic gesture should be extended to incorporate this phenomenon.

1.3. Aims

The aim of this research is to devise a novel approach for recognising a set of human gestures by analysis of a video recording from a single camera view. The recording would be made by a single camera directly in front of the gesturer who would be seated or standing.

An objective of the investigation was to confirm that the gesture trajectory could be modelled as an aperiodic waveform in 2D space and time. This modelling would allow Fourier analysis to be undertaken to show that the waveform could be characterised by its harmonic components. The normalisation of the harmonic series would allow gesture trajectory space and translation invariance from sequence to sequence.

An additional objective was the description of a 2D spatial trajectory using a complex representation of the waveform akin to 2D Fourier Descriptors, as a spatial-temporal, as opposed to purely spatial, signal.

It was also desirable to formulate a simple, robust method of detecting hand location for generating trajectory coordinates. Observation of a single gesturer in an image suggested that there were just three skin–coloured moving regions in an image: the face, and the two hands. An objective was formulated to merge skin-coloured cues with motion cues to form a combined skin-colour and motion cue that pinpointed the region of the hand in motion. An algorithm would then be developed to track the hand to provide the gesture trajectory.

Further objectives were set to investigate methods for classifying and recognising gesture trajectories from the harmonic components. These techniques were centred on types of neural networks and clustering methodologies that would be suitable for sparse experimental data.

1.4. Thesis content and organisation

The next chapter continues with an overview of gesture recognition. The definitions and classification of gestures, their dichotomies and typologies are considered. Reviews of the various approaches that have been taken for gesture recognition are made. The complementary work on avatars is discussed as to its usefulness in gesture analysis. The need for ground truth data with which to judge and compare the performance of recognition systems is discussed.

Chapter three details a reliable method for automatically generating data for the Fourier analysis described in chapter 5. This chapter investigates the use of colour and motion to capture hand location so that hand trajectory data can be obtained automatically. Additionally, methods of motion detection and background updating are investigated. The advantages of fusing colour and motion cues together are illustrated. Furthermore, the advantages of the unique rank ordering system of the skin-coloured and motion objects are explained.

Tracking of the hand position from the generation of skin-coloured, motion objects is discussed in chapter four. Environmental conditions and image sequence variables are experimented with, to gauge their impact on the quality of the tracking data. A description of the algorithm for object selection is explained for its use in correcting possible tracking errors in single-person; multi-person and two-handed sequences. The preparation of the data for the multi-rate normalisation technique is explained to overcome the different gesture lengths or sample rates.

Chapter five is the pivotal chapter of the thesis. It explains how a gesture can be considered as an aperiodic waveform in two dimensions of space and one dimension of time. The equations derived from object recognition using the Fourier Descriptor technique are modelled in the time domain instead of the spatial domain. The transformation of the time domain data into frequency domain data by standard Fourier analysis techniques derives positive sequence and negative sequence to characterise the gesture trajectory. The advantage of using exponential positive and negative sequence components is that the modelling (synthesis) of trajectories is in the same form as for the analysis of trajectories. The added benefit of working in the frequency domain is that through simple manipulation of the frequency data, the advantages of position and size invariance are obtained. The consideration of

individual harmonics gives an insight into the modelling of the trajectory as an infinite set of harmonically related 'elliptical corkscrews'.

This chapter also focuses on the invariance of the orientation angle. It is shown to characterise a gesture and its spatial characteristics. The orientation angle is also shown to be invariant to truncation errors in the gesture trajectory. Modelling the trajectory by just the first three harmonics and their associated elliptical characteristics shows a close relationship to the original trajectory data.

The advantages and characteristics of using RBF (Radial Basis Function) as a basic building block of PNN (Probabilistic Neural Network) are discussed in the following chapter, chapter six. The PNN is ideal for use in recognition activities when data is sparse but requires virtually no training. The key aspect of using PNNs is in the selection of target gestures. It is shown that clustering techniques can be a help in finding target gestures when gestures are indistinct. The hierarchical clustering method was used in this investigation because it was more suited to the sparse data than other clustering techniques. Various distance metric and linkage methods available for the clustering technique are considered. The clustering results are used to realise suitable target gestures for the PNN network. The performance of the PNN with pointing gestures is also considered.

Chapter seven discusses a further range of experiments that have been undertaken. The experiments considered gestures using a repeated gesture by the same gesturer; a gesture repeated by several people and a range of gesture stimuli interpreted by several people. The investigations show that with appropriate clustering tools and procedures with the PNN, inter-class and intra-class variations can be found. The results of gesture stimuli experimentation expose the oscillating nature of this type of action and a proposed refinement of the tri-phasic gesture definition. Observation of the response of gesturers to some of the experiments shows that skin-colour change and non-dominant hand movement can be an additional indicator of gesturer's internal, emotional state.

The final chapter reviews the thesis. It concludes with a review of the original aims and objectives, to reflect on the potential direction of future research in this subject.

2. Review of Hand Gesture Analysis and Gesturing Systems

This chapter clarifies the difference between static and dynamic hand gesture. It also reviews some of the definitions used to classify dynamic hand/arm gestures. Of the many different definitions in use the terms of 'bi-phasic', tri-phasic' and 'beats' need explanation as they are used by a number of vision researchers. Much of the impetus for gesture research is considered as coming from requirements of the signing community and from research into designing naturally performing avatars. The challenges of recording and analysing gesture actions are discussed and a review of recognition techniques is conducted. Of the many recognition techniques available, state-based techniques have been of most interest, notably Hidden Markov Models. The importance of verifying any recognition system by comparing recognition techniques and sources of 'ground-truth' data are also discussed.

2.1. Introduction

Human beings are adept at analysing gestures. Johansson's (1973) and Bobick's (1997) work in this area clearly demonstrate this. Johansson's famous experiments with just a few lights in motion against a dark background showed how sparse information was sufficient for an observer to detect enough to recognise human behaviour. Bobick showed a few frames, of extremely low resolution, of a subject performing a normally trivial recognisable movement. Although there was a lack of recognisable features in the still images, the movement is easily recognised when the still images are sequenced to suggest motion. Additionally, Kennaway (2001) reports that when the characteristics of one signer were implemented on a different avatar body the signer was still easily recognisable. Each of these examples indicates the power in the human system to recognise the unique characteristics of a particular signer.

Gesture recognition techniques have advanced in the last decade. This chapter reviews hand/arm gestures, their classification and analysis. It is found that the term 'hand-gesture' can be misconstrued. Sometimes 'hand-gesture' refers to static hand poses of just the hand but at other times refers to the dynamics of the hand and arm in movement. Freeman and Roth (1995) made the distinction between dynamic and static gestures.

"A static gesture is a particular hand configuration and pose represented by a single image. A dynamic gesture is a moving image represented by a sequence of images."

Static hand gesture tends to refer to hand shape and orientation. Whereas a dynamic hand gesture tends to be about the movement of the hand through space in which the hand shape is not so relevant. Wilson and Bobick (1995) make a similar observation that in some cases the spatial configuration of the hand is important, or alternatively the gross motion of the hand may be important. Quek (1994) also reported that it was rare for both the pose and the position of the hand to simultaneously change in a meaningful way during a gesture. Starner and Pentland (1995) recognise that tracking of the hand does not require a fine-grain description of the hand shape; studies have shown that such detailed information may not be necessary for humans to interpret sign language. Kennaway (2001) makes a comment about the amount of precision required in signing: -

"People learning to sign, learn from example and identify which parts of the action are significant and how much precision is required for good signing."

More recent work in classification of hand/arm gesture has explored in more detail the orientation pose and movement of the hand and arm structure. Some of this work has been initiated by the surge in interest of virtual humans or avatars, most typically work explained by Kennaway (2001). The many experiments in the synthesis of avatar design, complements gesture analysis. This is a source of accurate ground truth data for the movement of hand and arm during gesturing.

The goal of any machine system is to replicate as closely as possible the human system. To design a machine that can have these capabilities it is vital to have a better understanding of gesture, and then investigate methods to construct a system to detect and recognise human behaviour. But it has been found difficult to compare systems because of the lack of publicly available image sequences; different environmental conditions; and constraints imposed on the sequences.

2.2. Gestural Analysis and Characteristics

Consideration of what characteristics enable a gesture to be interpreted must be considered. The survey of Pavlovic et al., (1997) considers Gesture Modelling, Gesture Analysis, Gesture Recognition and Gesture-Based Systems and Applications. The paper extensively reviews the research activities that have occurred up to the published date and is biased toward hand gesturing rather than facial gesturing. The paper reports that hand movements can range from the simple action of using the hand to point and move objects around, to the more complex that express feeling, to actions to aid us in the communications of meaning with each other. The recognition of gestures require static and dynamic configuration of the human hand, arm and body to be measured. Cumbersome interface tools can only be justified in specialised application domains such as 'simulation surgery'. 'Glove-based' devices are reviewed by Baudel and Baudouin-Lafon (1993) and Sturman and Zeltzer (1994). They concur that 'glove-based' devices will ultimately deter everyday users from using such a system.

Psychological studies have reported useful facts about human gesture. Kenden (1986) states that there is 'autonomous gesture' and 'gesticulation', the former gesture acting independent of speech and the latter is in association with speech. He shows that the temporal nature of hand gesture can be divided into three parts; preparation; nucleus (peak or stroke) and retraction. The preparation phase consists of a preparatory phase that sets the hand in motion from a resting position. The nucleus phase has a definite form and enhanced dynamic qualities. The hand then returns to the resting phase i.e. retraction, ready for the next gesture. The exception to this is the characteristic of 'beats', an alternative type of gesture that is related to the rhythmic structure of speech.

McNeil and Levy (1982) found that as well as 'beats', the type of action inferred by a gesture could be categorised as iconic, metaphoric and deictic gestures. Iconic gestures are air pictures used in the mime game, charades to indicate the categories of a 'book', or 'film', or 'play'. Metaphoric gestures are used to represent abstract

concepts, for example the pinching action to indicate precision. Deictic gestures are pointing actions to indicate position or placement. 'Beats' are like small hand waves that simultaneously combine with vocal inclination, to stress or emphasise parts of the speech that are significant or important.

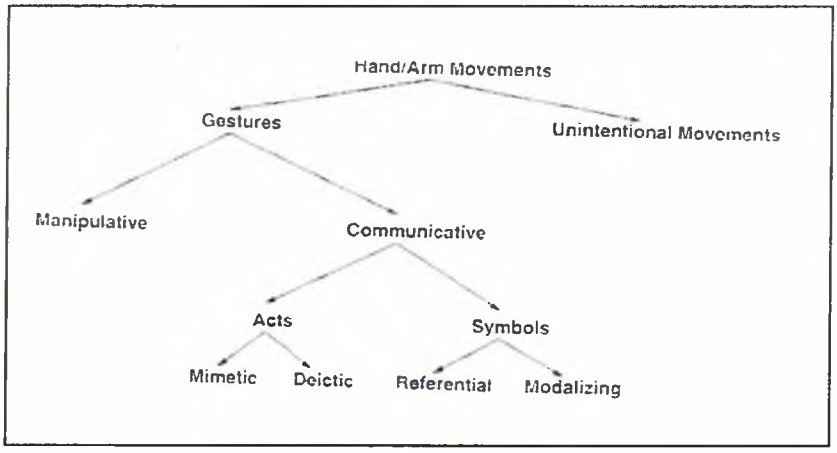


Figure 2.1 Taxonomy of Hand/Arm Movements (Source Pavlovic et al., 1997)

A general taxonomy of Hand/Arm movement is presented by Pavlovic et al, (1997), attributable to original work by Quek (1994) as shown in Figure 2.1. The movements are sub-divided into different classes.

Wilson et al. (1996) characterise the iconic, deictic, metaphoric and beats by their temporal signatures. Each gesture is bracketed by 'rest states'. The simplest gesture, 'beats' consist of small baton-like movements away from the rest state and then back again. This type of gesture is termed 'bi-phasic' whereas the other gestures are termed 'tri-phasic' because of the three distinct phases of transitioning from the rest state to gesture space, executing the stroke phase and then transitioning back to the rest state.

Other significant work on gesture classification was undertaken by Ekman and Friesn (1969) and Argyle (1975). Ekman and Friesen divide communicative gesture into four characteristics of 'Emblem' (acts which can be translated into words); 'Illustrators' (movements that illustrate verbally expressed content; 'Affect' displays (facial expressions displaying the speaker's feeling; and 'Regulators' (semi-conscious acts regulating the speech flow rhythm. 'Adaptive' gestures are also categorised. This refers to non-communicative gestures, and are unintentional but they respond to physical need. Likewise Argyle categorises communicative gesture into the conventional gesture whose meaning can be interpreted into words and speech related gesture in which movements underlay the meaning conveyed by the speech flow. Argyle also differentiates non-communicative displays into those that have no communications function and the idiosyncratic gestures that reveal the gesturer's personality, mood or behaviour.

Furthermore, Rossini (2003), attempts to define interdisciplinary parameters for gesture analysis. This analysis defines four components: gesture size (angle of moving part of joint, with respect to the horizontal, during the stroke phase); gesture timing (the gesture phase between the pre-stroke phase and retraction); point of articulation (main joint in the gesture movement) and locus. The locus defines the personal space or gesture space that is used. It is analysed by classical Cartesian axes. The 'x' axis is divided up into left and right peripheries. With periphery 1 taken

up by the distance of the elbow from the trunk and periphery 2 taken with the hand fully extended. The 'y' axis is divided into the overhead, head, upper bust and lower bust space. The 'z' axis is sub-divided into trunk, middle distance and full distance space. The gesture size component was particularly helpful in clarifying mobility patterns in deaf signing people.

In comparison, Pavlovic et al. (1997) consider that a gesture is a stochastic process in the gesture model parameter space over a suitably defined time interval. The importance of this definition is that no two realisations of the same gesture will result in exactly the same hand/arm motion or the same set of images. The presence of the time interval emphasises the dynamic nature of the gesture. Although gestural activity is essentially spatial, the temporal content is of significance. The gesture classification technique must be both time instant invariant and time scale invariant. The example given is that of clapping. The gesture should be recognised whether it is performed slowly or quickly. Automatic speech recognition deals with such problems. The recognition of the spoken word is independent of their duration and variation in pronunciation. In practice the variations of spatial and temporal variations can be challenging in hearing impaired people. Conversation with a Makaton specialist explained that there can be a problem of interpreting the signing of impaired people. It was clear that impaired people have a large variation in their gesture action, which sometimes is difficult even for a trained human observer to understand.

Fundamentally there appears to be some difference of perception between the definitions and characteristics of gesture relating to natural gesture and signed gesture. One observation is that natural gesture is inherently speaker dependent, influenced by cultural, educational and situation factors (Wilson et al., 1996). They relate to the inter-class variations of gesture. Most of the work on gesture interpretation has been focused on the understanding of sign language. However, gesturing is a part of 'normal' peoples communications experience and adds benefit to the communications process. The variability of human behaviour (Wilson and Bobick, 1995) of the gesturer must be described without regards to precise geometry or precise temporal information. We take visual behaviour to mean the sequence of visual events that make a complete action or gesture. Sign language (ASL) has its own grammar rather than borrowing grammar from English (Starner and Pentland (1995). This grammar allows more flexibility in word placement and sometimes uses redundancy for emphasis. A signer may describe a person, place or thing but then point to a place in space to temporarily store that object for later reference. In addition the position of the eyebrows is used to indicate question, statement or directive.

Further work on the characteristics of gesture has been made through the development of avatars or manikins. In order to design avatars, greater understanding of the structure of hand/arm movements is necessary (Gibet et al., 2001). The study of sign languages has highlighted the convergence, not only of the linguistic features but also in the formational and functional parameters characterising gesture. It has been observed that the French Sign Language (FSL) gestures have five co-occurring 'parameters': -

- The configuration that is the hand shape
- The orientation that gives the directions pointed by the palm and metacarpus

- The movement which is generally a description of the arm's endpoint kinematics
- The location, which is the area where the sign occurs. A same sign may be indeed realised in different parts of space, depending on its meaning
- The facial expression, which has a complimentary role in the sentence by giving mode, for example.

The disadvantage of the motion capture technique, for avatar design, as described in chapter one, is the large amounts of data that is generated, although the data has great accuracy and resolution. The alternative method is to use a simplified biomechanical model for the particular application of signing for a real-time animation. The HamNoSys (Prillwitz et al., 1989) was developed to transcribe signing gestures. Each gesture is broken down into components such as hand position, hand orientation, hand shape, motion etc. But the system has been designed to be read by people rather than computers.

More recently HamNoSys has been upgraded to version 4 (Kennaway, 2003) to include substantial coverage of facial expression because speech-like movements of the mouth are frequently used in signing. The symbolic language is transformed by a program and conveyed to the avatar. To aid computer animation the syntax of HamNoSys is translated to SiGML (Signing Gesture Markup Language). Another program then takes the SiGML representation of a gesture to add default location, timing durations, and rotation of each joint to produce appropriate animation data. However, movement speed descriptors are limited to the categories of merely fast, slow or ordinary speeds. Finally, the avatar is rendered to display the avatar on the screen at the specified times, in the specified postures (Appendix I).

The avatar is defined by the position of all the joints (Appendix I) when placed in some standard pose. Information is also required of each joint and whether it operates as a hinge, a ball and socket. Information is also needed whether there are any limits of movement. Defining gesture can be a complex task as there are 12 standard hand shapes with a set of modifications that can be applied to them to bend individual fingers or thumb. Kennaway (2003) says that the position can be specified in several hundred locations with 26 possible orientations: -

"Movement descriptions can be quite complex. A movement of the hand through space can be straight (in any of 26 directions), curved (the plane of the curve being orientated in 8 different ways about the axis of movements also quite complex with the hand being able to move through space in many directions."

Lifelike gestures can be enhanced by the addition of 'ambient motion'. Ambient motion is the small, random motions of the head, eyes and torso that are prevalent in human gesturing, and makes the animation more natural. It is recognised that synthesised data can be too good. Howell et al. (2003) note that the hand trajectory data was highly accurate, from the magnetic sensors, and applied random data so as to simulate less constrained data which might be extracted from visual methods.

2.3. Gesture Recognition Techniques

Gesture tracking can be considered as a sub set of human movement. Gavrila, (1999) made a detailed analysis of human movement, particularly whole body and hand movement. Gavrila also overviewed a number of developments in this domain and identifying a number promising application. Similarly, Pavlovic et al. (1997), reviewed Hand Gesture Modelling, Analysis, and Synthesis. These papers represent a very thorough review of the subject area. The tracking of the hand, in a dynamic gesture, has to contend with the spatial and temporal variability of the gesturer, the position and the image size.

The most successful systems have been centred on state space techniques to overcome these variabilities. The state-space techniques mostly use Hidden Markov Models (HMMs) although there are exceptions, notably Bobick and Wilson (1995, 1997). Other techniques that do not use space techniques are Temporal Templates (Davis and Bobick, 1996), Motion History Images (Bobick and Davis, 2001), Neural Networks (Howell and Buxton, 1998). Darrell and Pentland (1993) use dynamic time warping and normalised correlation to match the interpolated responses of several learned image templates. Input to the HMM models use a suitable set of features. This can be the spatial coordinates of the hand, with perhaps some other attributes of the hand shape (Starner and Pentland, 1998). Pavlovic et al. (1997) report that features can range from region based parameters, like colour and motion; wire mesh; orientation histograms; or facets of whole images given by eigenvector coefficients, PCA or Gabor wavelets.

There are many different ways to analyse gestures. Kohler (2001) compiled a table of vision based hand gesture recognition systems. Comparisons are made between the task, segmentation, features and classification technique. Interestingly only some fifteen systems used colour segmentation out of the forty or so techniques listed. The features are very varied, ranging from, fingertip detection; edge detection; centroid mass; shape moments; boundary tracking; eigenspace; Fourier descriptors; moments of difference images; orientation histograms; 2D/3D point distribution model; regions and blobs; silhouette; coarse direction and magnitude values; Gabor filter and Zernike moments. The classification techniques range from neural networks (NN); Multi Layer Perceptron Neural Network (MLP NN); Hidden Markov Model (HMM); continuous HMM; coupled continuous HMM; correlation, local shape property learning; hand/non-hand classification; finite state machine, distance from feature space; 3D cylindrical finger model; k-means for active shape model (2D smart snake); inverse kinematic model; stochastic deformable model; multiple classification; 3D hand skeleton model; finite state estimation; convolution, template matching.

Originally, the burden of analysis by gesture recognition techniques was lowered by the use of passive or active markers or marked gloves (Huang and Pavlovic, 1995). Others use restrictive set-ups: uniform background, limited gesture vocabulary or just a simple static posture analysis. Very few techniques today use markers, but restrictions or constraints are still prevalent, indicating the complexity of the recognition task in real environments.

2.3.1. Spatial Modelling

Pavlovic et al. (1997) explain that Spatial Modelling of gesture can take various forms depending on the application, and can be broadly sub-divided into the two classes of 'appearance-based' models and 3D models. The 'appearance-based' model is where gesture is revealed directly from the images observed. The 3D approach is where a gesture is inferred from the model of the motion and the position of the hand. 3D Hand/Arm models can be broadly classified into either volumetric (cylindrical) or skeletal models. The types of modelling of the hand can be very varied i.e. 3D textured volumetric model; a 3D wire-frame volumetric model; a 3D skeletal model; or a binary silhouette or contour model.

Appearance-based models of the hand and arms model the appearance of the gesture to a predefined template of gestures. There is a large variety of models in this group. Some use deformable templates of hand, arms and bodies, or trajectories. Gavrilla (1996) compares the number of techniques used in 2D (with and without explicit shape models) and 3D approaches. There were twenty-six different 3D approaches catalogued by Gavrilla (1999) in the paper 'Visual Analysis of Human Movement'. There was almost the same number (twenty-three) of techniques on 2D approaches without explicit shape models, whereas only thirteen approaches using explicit shape models were given.

In addition, sequences are used as gesture templates or temporal templates (Davis and Bobick, 1996). A gesture is modelled by a sequence of representative image n-tuples. However, Sherrah and Gong (2001) showed that rather than relying on spatial-temporal continuity and complex 3D models of the human body, a Bayesian Belief Network could be used to deduce the body part positions by fusing colour, motion and coarse intensity measurements with contextual semantics.

2.3.2. State-Based Modelling

Hidden Markov Models have been very successfully applied to a range of applications especially speech recognition (Rabiner, 1989). The use of Hidden Markov Models (Starner and Pentland, 1995), (Yamato et al., 1992) and (Chen et al, 2003), has been transferred into gesture recognition. HMMs consist of a number of states which can capture the underlying nature of the gesture from a set of training examples. This approach overcomes the problem of variability, uncertainty and probabilistic nature of gesture (Sage et al, 2003). HMMs are further described by the set of probabilities that a state gives rise to and the probabilistic transitions between states. Many of the examples that use HMMs are for the analysis of human dynamics that show a continuous or periodic nature. However, Kobayashi and Haruyama (1997) argue that HMMs are not necessarily appropriate for modelling gesture features that are transient. They proposed the partly hidden Markov model that showed seventy-three percent improvements in error rate over HMMs for isolated sign recognition, although these results were for a restricted set of just six signs.

Alon et al. (2003) proposed a technique for clustering time-series data to discover groupings of similar object motions that were observed in a video collection. A finite mixture of HMMs is fitted to the motion data using the expectation-maximisation framework. The formulation allows each sequence to belong to more than a single

HMM. The decision about class membership can be deferred until a later time when such a decision is required. Promising potential for this technique is claimed on a number of experiments including camera mouse experiments and gait experiments. In the latter experiment, the input time-series is modelled as sine waves plus noise. A periodic structure is defined as consisting of four states corresponding to the sine valley, zero crossing up, peak and zero crossing down. The typical model consists of as many states as observations in a single period. The classification accuracies of the clustering algorithms were very similar to the classification accuracies obtained with supervised learning i.e. without the need for class labelling.

Starner and Pentland (1995) used a feature vector of eight elements taken from the blob of each hand representing 'x' and 'y' position, angle of axis of least inertia, and eccentricity of the bounding eclipse. The axis of least inertia is determined by the major axis of the bounding ellipse, but this can lead to a 180-degree ambiguity in the angle of the ellipse. However, this problem was successfully addressed by allowing the angles to only range from –90 to +90 degrees.

The earlier work of Yamato et al. (1992) explores how the HMM technique is applied to recognising six different tennis strokes from 200 by 200 pixel image sequence. The model parameter approach of the human form is dismissed because it is not robust or reliable for real images. Low-level image features such as the area of the subject are justified, as they are more robust than model fitting procedures. The results gave over a 90% recognition rate when the training data and test data was of the same subjects. But when the training data and test data was from different subjects the recognition rate fell. The performance depended on the number of training patterns used and how well the patterns were representative of the spread of category. It was recognised that people had some unique quality in their action. Further work is required on refining feature extraction.

Bauer and Kraiss (2001) introduced an HMM-based continuous sign language recognition system using subunits. Signs of the vocabulary are made up by the concatenation of the sub-units. This aids the enlargement of the vocabulary by reducing the training material. The feature extraction was aided by the use of a coloured glove. Recognition rates achieved were about 80% with 12 different signs and 10 sub-units. Cheng et al. (2003) produced a hand gesture recognition system using a real-time tracking method and hidden Markov models. The significance of this work is that the input feature to the HMMs were not coloured gloves but were a feature based on skin-colour and motion. The system was tested to recognise twenty different gestures achieving a recognition rate of 90%. The performance was affected by some of the signers imprecision in signing. In addition some of the error was because there was insufficient training data to make a good estimate of HMM model parameters from the 1200 sequences representing the twenty gestures.

Simple gestures for visually mediated interaction, using motion image moments, have been investigated by McKenna and Gong (1998). Four deictic gestures were studied for motion of the arm and hand where the shape of the hand is unimportant. Trajectories based on a simple set of motion image features were used to estimate models for gesture events. The feature trajectories were appropriately time-scaled. Recognition was performed using a Gaussian matching function. This function, for gesture recognition, used a probabilistic finite state machine similar to a HMM, although the transition probability matrix is not used. Results were acceptable

although there was some confusion between high waves and low waves. It was also noticed that the subject was sitting relatively motionless making no major changes during the image sequence.

The technique of Bobick and Wilson (1997) defines a gesture to be a sequence of states in measurement or configuration space. In this work, a training set of trajectories was used to capture both the repeatability and variability of the given gesture. A prototype trajectory was generated from an ensemble of trajectories. Configuration states were developed from the prototype. Gesture recognition was undertaken from an unsegmented, continuous stream of data from two dimensional movements of a mouse input device; hand movement from a magnetic spatial position and orientation sensor and eigenvector projection coefficients computed from an image sequence. The second experiment using a magnetic sensor yielded somewhat sparse, higher dimensional data; so as to ensure that there were enough points available to compute the prototype curve, each example was up-sampled using splines, to give each gesture forty samples and each point gesture seventy samples. This technique represents gesture by a time-invariant but order-preserving method based on a convenient arc length parametisation of the data points to produce the sequence of states. There is similarity in this work with HMMs, namely the existence of states. However, the important distinction is the production of a prototype, which is easily and readily available. The HMM, in contrast is timely in production because of its statistical nature with the adjustment of many free parameters. The paper does question the validity of the approach, in that the spatial configuration is the most important aspect of the signal to be extracted. It is suggested that the temporal properties may be the more important elements of a gesture but to consider the temporal structure or gestural phases in natural gesture. The authors also suggest that the technique is appropriate for stylised or literal gesture, but inappropriate for a natural gesture or the spontaneous gesture generated by a person telling a story for instance. The paper concludes that there is little consensus in the literature on a useful definition of gesture. The development of the state technique attempted to formalise a notation of gesture that was not limited to a particular domain, as illustrated by the three very different experiments.

Recent work on real-time hand tracking (Stedfanov et al., 2005) uses Variable-Length Markov Models (VLMM). Automatically acquired VLMMs are used for tracking of structured behaviour and are used to represent high-level structure and also temporal ordering of gestures. This work demonstrated that the approach combines behavioural knowledge with a stochastic simulation to achieve robust tracking of hands in a human-computer interaction environment. The task of manipulating virtual objects in a natural manner, using hand movements and gestures did cause some problems for the users. The authors suggest that gestures must be carefully designed so that users can employ them comfortably and straightforwardly to signify different actions. The results indicate that what seems natural may not, in fact, be the most effective design. This warrants further investigation and explanation.

2.3.3. Alternative Modelling Techniques

A different approach to gesture recognition is based on time-delay invariant Radial Basis Function (RBF) network (Howell and Buxton, 1998). Gesture motion is

detected by differencing consecutive frames and a sparse arrangement of Gabor filters is used to process the differenced images. This information is applied to a TDNN (Time-Delay Neural Network), using one RBF network for each training example. Only simple, limited techniques for the temporal segmentation of gestures are used but high levels of performance can be obtained. The main disadvantage reported was the difficulty in classifying the same behaviour at different speeds using a single time window. Another vision-based system, Ng (2002), for interpreting hand gestures used a procedure to extract binary blob(s) of the hand. The shape of the blob was represented by Fourier descriptors input to a RBF for pose classification. Further processing with HMMs and Recurrent Neural Networks (RNN) was used for the recognition phase.

Yang et al. (2002) employed TDNNs to classify motion patterns of hand regions as a consequence of them being successfully applied to the spatial-temporal patterns in phoneme recognition (Waibel et al., 1989). Waibel et al. demonstrated that using the TDNN gave lower error rates than that achieved by a simple HMM recogniser. It is noted that a TDNN has two important properties. First, it is able to recognise patterns from poorly aligned training examples, which is to be expected from the slight difference in duration of gestures. Secondly, the total number of weights in the network is relatively small since only a small window of the input pattern is fed to the TDNN at any instance. Another important feature is the temporal integration at the output layer that makes the network shift invariant i.e. insensitive to the exact positioning of the gesture. The input feature vector consisted of position coordinates; velocity magnitude and angle values. The parameters in the TDNN (number of nodes in each layer, number of hidden layers and window size) were selected empirically by numerous experiments on a training set.

2.4. Ground Truth Data and Comparisons

Reviews of many papers on gesturing have shown different success rates of the varying techniques in use. The paper by Lockton and Fitzgibbon (2002) examines gesture recognition for single-handed gestures (static) using a 'deterministic boosting' algorithm that does not use a temporal Markov model, with a 99.87% success rate. However, this was only achieved through a wristband to replace accurate lighting control. The lighting was the same for the test and training data. Likewise Kohler's (2001) table of comparing vision-based hand gesture systems also comments on the constraints (user dependent/independent, static background, background independent or optimal illumination), speed, trained gestures and recognition and error rates.

Further work on comparing the claims of different techniques was investigated by Morrison and McKenna (2003) by comparing Trajectory-based and History-based representation techniques. A direct experimental comparison of these two approaches is presented using skin colour as a common visual cue, using recognition methods based on HMMs, moment features and normalised template matching. The constraint imposed on the recognition system was that it must be capable of learning gesture models from only a few examples, as users find the process quite tedious after about ten examples per gesture. Importantly, the comparisons were simplified by ignoring the temporal segmentation problem and used isolated gestures. The comparison was tested on three one-handed and three two-handed gestures. It was interesting to note

that the gesturers were allowed to specify their own gesture vocabulary so that they would be more comfortable performing and remembering the gestures. The history-based approach achieved consistently higher recognition accuracy using template matching than when moment features were used. The best in terms of computation and error rate when the template matching was based on mean absolute difference (MAD). The trajectory-based approach using HMMs gave better results than moments with Skin History Image (SHI), but did not perform as well on average as template matching of SHIs. The differences in recognition rate between these techniques was not significantly higher as trajectories with HMMs gave 82.7%, SHIs with MAD gave 89.3%, SHIs with mean squared difference (MSD) gave 87% and SHIs with correlation coefficient (CC) gave 85.2%. However, the use of central moments (CM), scale-normalised moments (S-nom) and HU-moments only gave recognition rates of 67.3%, 62% and 41.8% respectively. The analysis of the errors made using HMMs and MAD SHIs showed that the two approaches made different errors. This would suggest that it might be possible to combine the approaches to further reduce error rates.

A method of performance evaluation of trajectory detection is to provide ground-truth data with which to compare experimental results. Black et al (2003) report on a number of semi-automatic tools that are available for generating ground truth for video surveillance tracking systems. Whether automatic or not, a system will provide independent and objective data (e.g. classification, location and size) that can be related to data from the video sequence. The gathering of manual/visual ground truth is usually undertaken by a human operator who 'points and clicks' at the frames in a sequence, at well defined points of interest. The resulting trajectory of points is then compared with the points generated from the tracking performance.

In order that the performance can be evaluated on a 'level playing field', image sequences are now becoming available for test purposes. In the past each researcher needed to generate their own image sequence, resulting in a variety of environmental and lighting conditions. This problem has been confronted by PETS-ICVS Datasets (2003) and FGnet (2004) database. The PETS-ICVS consists of datasets for a smart meeting. The environment consists of three cameras: one mounted on each of two opposing walls, and an omni-directional camera positioned at the centre of the room. The overall task is to automatically annotate the smart meeting (Appendix VII). FGnet has also available many databases for example pointing and command gestures under mixed illumination.

Another possible set of ground-truth data is available from the avatar designing community. The design of the avatar can be made from two possible sources, as mentioned in Chapter 1. The position of every joint is held by three values to specify Translation and four values to specify Rotation (Quaternion Notation, Appendix I). For example, to find where the wrist is, you need all the translations and rotations at the joints up the hierarchy from the wrist to the root need to be combined. The testing of gesture recognition routines can be made on the data files that hold the avatar animation sequence. At present, some of these sequences are not as complex as real sequences and some attempt is made to add 'ambient motion' to make a gesture more lifelike. It is expected that progress will be made to make the whole image more lifelike in the future.

2.5. Summary

The definition of gesture continues to be refined. Psychologists have a similar terminology for describing gesture characteristics, but there is no universal agreement. There can be some confusion with using the word 'hand-gesture' too loosely. In some instances it means the static pose of the hand whereas in other situations it means the dynamic trajectory of the hand in gesture-space. The taxonomy of hand/arm movements seems to be a helpful approach to categorising the movement into its many uses and meaning, as does the terms bi-phasic and tri-phasic as an indication of the type of gesture structure and activity. However, more recent work on gesture has concentrated on a detailed description of spatial structure. It has been found that gestures have to be specified with many parameters. The advent of avatar design has shown that virtually all joint movements and positions need to be specified to enable typical human gesture characteristics to be observed or modelled.

Gesture analysis originally used glove-based data, but then moved on to using natural features with appearance-based models and 3D models using a variety of classification techniques. HMMs have shown great promise in the classification process because of their time invariant characteristics. However, one of the biggest problems is the training required to make the system work with little error and its inflexibility at expanding the vocabulary. Chen et al. (2003) attributed some of the errors to insufficient training. Correspondingly Morrison and McKenna (2003) note that users find giving more than ten examples tedious. Other constraints in systems tested relate to non-complex backgrounds; users sitting quite motionless in the context of non-variable lighting conditions.

HMMs are not the only unique method of describing and classifying gestures and many other techniques have shown promise. Morrison and McKenna compared HMMs with SHI techniques with isolated or segmented gestures. This showed that the latter technique gave better results although there were not large differences in performance when template matching was used with the SHIs. Howell and Buxton (1998) report that the use of RBF and TDNN resulted in very good performance compared with other techniques like, for example 'moments'. Chen et al. (2003) approaches the use of feature cues in a similar way to this thesis by fusing of motion and skin-colour cues. Fourier Descriptors techniques are also used but for hand shape recognition. The recognition technique uses HMMs rather than the PNN described later in this thesis.

Recent work in the gesture community has focused on using common data sets to compare the effectiveness of techniques. In this area the 'ground-truth' data is seen as another important aspect at verifying experimental work. The comparison and effectiveness of techniques has been undertaken by some researchers. Bobick and Wilson (1997) undertook some experiments on a state based approach which used three different input sources. A significant comparison of techniques has also been made by Morrison and McKenna (2003) in which HMMs and SHI where compared using the same feature cue. Interestingly, there are few comparisons in the literature on comparison of techniques. The advent of common data through organisations such as PETS and FGnet should produce more comparison material in the future.

The next chapter discusses a technique of fusing skin-colour and motion cues for tracking the hand position in its gesture trajectory. The technique is tested in a range

of lighting environments. A suitable colour-space model is applied to skin-colour detection. The technique is also modified for use in a range of lighting conditions due to the environment or due to incorrect white balance. Fundamentally the skin-colour cue is linked to a motion detecting cue to produce a reliable moving skin-colour object, referred to as a gesture object. The rank ordering by area of the gesture objects is shown to be a very effective aid at detecting the gesturing hand and its associated trajectory.

3. Detecting Hand Position by Colour and Motion

The development of how a simple and robust method of detecting hand position of a gesturer is explained in this chapter. Firstly, consideration was made of detecting the hand by skin-colour. In order for this to occur, suitable colour-space models were considered that were invariant to light intensity change. A review of many of the colour-space models was made, but the HSV version was found most appropriate for this application. Consideration of the 'H' (Hue) properties of this model showed that there was a discontinuity in the red region that could affect skin-colour detection. The solution to the problem is explained. The chapter then proceeds with detecting motion in the image sequence and proceeding to merge the skin-colour and motion cues. The merging of the cues produce skin-colour and motion objects which when rank ordered by size invariably locates the hand by the most significant object. A number of experiments and variations to the production of the skin-coloured objects is investigated for optimal conditions of segmenting the hand from the image, regardless of environmental conditions.

3.1. Introduction

The previous chapter reviewed the various features and classification techniques used to recognise gestures. Although gesturing can be viewed as a whole body experience of movement, in most cases gestures can be conveyed by the observation of the movement of just the hands and head. From this observation an hypothesis can be formed that in an image of one gesturer there are typically three areas of motion associated with skin colour i.e. the two hands and the face. Furthermore, with a single-handed gesture the motion of the dominant hand is the most important skin-coloured object to track. The motion of the head and less-dominant hand can often be ignored.

Motion in an image can be detected by comparing the difference in intensity in two adjacent frames (Jain et al, 1995). This assumes that lighting levels remain constant between frames, which is generally the case with gesturing as the duration of the gesture is a relatively short time of just a few seconds. Methods for adjusting for background variation are possible (Stauffer and Grimson, 1999) using adaptive background estimation. Adaptive techniques are generally found to add to the computational burden and affect the updating speed. At normal sequence capture rates of 25 frames per second the gesture duration is quite short and there are few samples that are obtained to capture a reliable history of pixel variability. Additionally, variations to pixel levels can be caused by automatic adjustments in the camera controls during the recording of an image sequence.

The segmentation of difference pictures, by a suitable threshold value, produces a binary mask with many different sizes of objects. The objects can result from various sources, not just skin-coloured motion, particularly in complex scenes. Many of the resulting objects can be considered as noise. Setting the threshold at an appropriate level can be a challenge for all possible image sequences. A new technique is introduced of sorting the objects into rank order by area. The largest movements are associated with the most significant size of objects and generally relate to the movement of the dominant hand. Heuristic approaches to setting the threshold level

are not required. As a result threshold levels do not need adjustments from sequence to sequence and image size to image size.

Rank ordering can be particularly effective when the motion cues and the colour cues are combined by the fusion of the segmented binary masks (Cheng et al., 2003, and Harding & Ellis, 2003). This technique significantly reduces the number of objects to process. Furthermore, size ordering these objects ensures that, in most cases, the most significant object relates to the movement of the dominant hand.

One particular problem associated with the recording of gestures is the clothing worn by the gesturer. In most sequences the gesturer wears a long sleeved shirt of some kind, of different colour to the hand and so significantly helps in the segmentation of the hand region. There are a number of sequences when the subject wears a short-sleeved shirt or blouse exposing a much larger area of skin coloured region. This results in either the centre of gravity of the skin-coloured region being significantly different to the hand region or the production of more than one object due to local variations of skin-colour or motion. This problem was investigated by Cheng et al. (2003) by detecting the regions of high 'edgeness' around the fingers using Kirsch template matching techniques (Vernon, 1991 and Davies 1997).

Testing has also been undertaken using publicly available test sequences. One of the organisations, PETS (Performance Evaluation of Tracking and Surveillance systems) has recently produced 'Smart Room' sequences of different scenarios. Another organisation to make available sequences for face and gesture recognition is the FGNet Network of Excellence in Face & Gesture Recognition. The paper by Holte and Storing (2002) documenting the sequences that FGnet publish, explains the problems that are encountered at present with the combination of artificial indoor light and outdoor illumination (through windows). It is noted that: -

"A cluttered background and such illumination conditions make the low level segmentation of computer vision-based gesture interfaces often fail. In particular, skin colour like objects and illumination colour changes are difficult to cope with, whereas the problem of high intensity ranges will be solved by future camera technologies ... that can capture much higher ranges than the human eye."

The importance of illumination levels and illuminants has been recognised (FGNet, 2002). An experiment for hand gesture recognition is described as to the environmental and cameras set up. The lighting is arranged so that a table (where hands are placed) is split in two parts with the same intensity (measured with a luxmeter). One side of the table has a colour temperature of 2600K and the other 4700K. The aperture of the camera is set to 2.2 and white balance has been performed with a colour temperature of 3400K and the camera is calibrated to have an offset/black current close to zero.

The work described in this chapter was instigated to automatically generate data for Fourier analysis This chapter explores a method used to reliably combine skin-colour and motion cues to track dynamic hand gestures over a range of lighting conditions.

3.2. Colour Fundamentals and Models

3.2.1. CIE Definitions

Colours are perceived in an image as a result of the spectral content of the ambient light, the spectral response of the sensors in the imaging system, and by the spectral reflections of scene surfaces. Although these three factors are complex the use of colour is motivated by it being a powerful object descriptor and the experience of the human vision system being able to discern thousands of colour shades and intensities compared to about two-dozen shades of grey (Gonzalez and Woods, 1992).

All colours are seen as variable combinations. The structure of the human eye has required colours to be viewed as variable combinations of the so-called 'primary colours' red (R), green (G) and blue (B). For the purpose of standardisation the CIE (Committee Internationale de l'Eclairage) in 1931, specified the three primary colours as blue = 435.8 nm, green = 546.1 nm, and red = 700 nm. It is important to note that no single colour can be called red, green or blue. In addition it does not mean that these three fixed RGB components acting alone can generate all the spectrum colours.

Colours can be distinguished from each other by reference to their brightness, hue and saturation. Hue is associated with the dominant wavelength and is the colour perceived by the observer. Saturation refers to the purity of the colour and refers to the amount of white added to the colour. Hue and saturation taken together are called chromaticity and therefore a colour can be characterised by its brightness and chromaticity. The amounts of red, green and blue to form any particular colour are called the tristimulus values and denoted, X, Y, and Z. A colour is then defined by its trichromatic coefficients and defined as: -

$$x = \frac{X}{X+Y+Z}$$

$$y = \frac{Y}{X+Y+Z}$$

and

$$z = \frac{Z}{X+Y+Z}$$

Hence, from these equations, $x+y+z=1$, or $z=1-x-y$

The tristimulus values needed to produce a colour corresponding to a particular wavelength can be obtained from curves or tables that have been experimentally obtained.

An alternative and common approach is the chromaticity diagram which shows the colour composition as a function of x (red) and y (green). The value of z can be determined from the equation above. The plot of the curve corresponds to monochromatic spectra. The shape has been variously described as 'shark-fin-shaped' (Sharma and Trussell, 1997) to 'tongue shaped' (Gonzalez and Woods,

1992), as seen in Figure 3.1. Pure colours are on the boundary of the locus and are completely saturated. Any point not on the boundary represents some mixture of spectrum colours. There is a point in the shape where there are equal fractions of the primary colour and represents the CIE standard for white light. As a point leaves the boundary and approaches the point of equal energy, more white light is added and it becomes less saturated.

A straight line drawn between any two points on this diagram defines all the colours that can be produced by the combination of the two points. To determine the range of colours that can be produced by three primary colours, lines are drawn between each point to produce a triangle. This triangle does not enclose the entire region of the 'shark-fin shape' and shows that not all colours can be obtained with three single primaries. This technique is often used to compare the relative gamut of RGB monitors and different printing inks. Figure 3.1 (Agfa,1994) shows the relative gamuts of RGB monitor and Pantone and SWOP-CMYK printing inks.

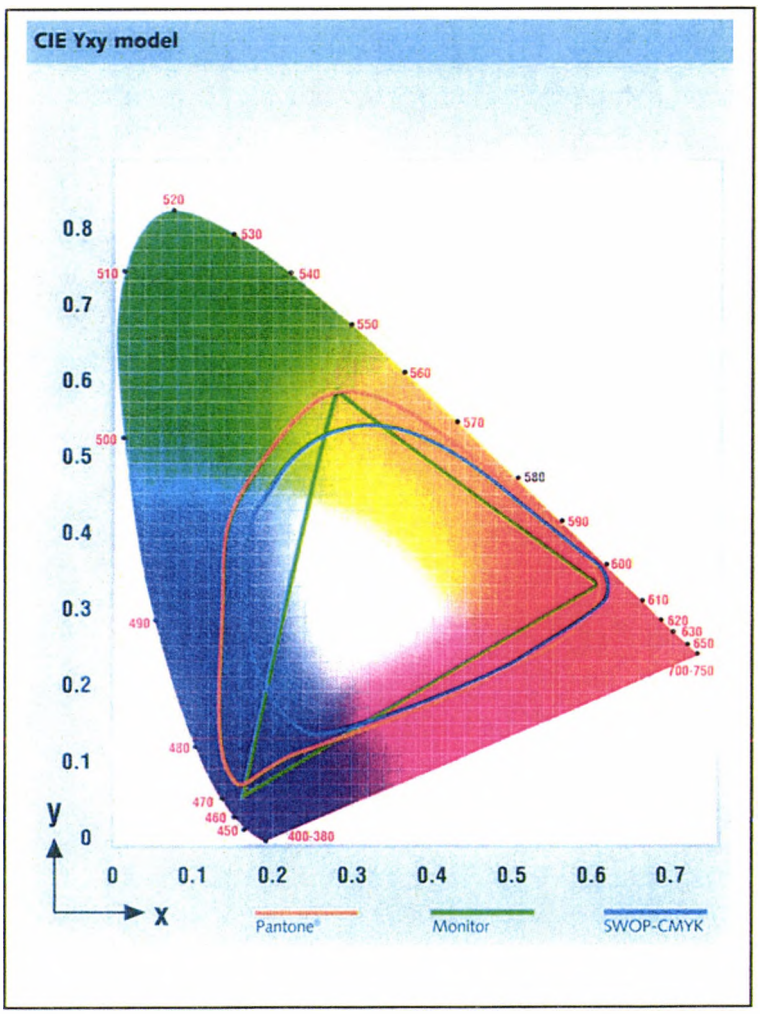


Figure 3.1 CIE Yxy model comparing Pantone, Monitor and SWOP-CMYK colour gamuts (source: Agfa)

The non-linear CIE Yxy colour model was mathematically transformed in 1976 to the uniform L*a*b* model, in which distances between colours more closely match

those perceived as shown in Figure 3.2 (Agfa, 1994). All colours of the same lightness lie on a circular flat plane, across which are the a* and b* axes. Positive values of a* are reddish, negative values of a* are greenish. Whereas, positive values of b* are yellowish and negative values of b* are bluish. Lightness varies in the vertical direction.

The colour models of XYZ, L*a*b* (CIELAB) and L*u*v* (CIELUV) all arise from colorimetry issues of modelling the human vision system and being able to match colours in different illuminants. Poynton (1997) explains that a perceptually uniform system is one that if a small perturbation to a component value is approximately equally perceptible across the range of that value. Both L*u*v* and L*a*b* improve the 80:1 or so perceptual nonuniformity of XYZ to about 6:1. However, both demand too much computation to accommodate real-time display, although both have been successfully applied to image coding for printing. It is noted that a computer vision system does not necessarily need to model the human eye to extract colour information for object identification.

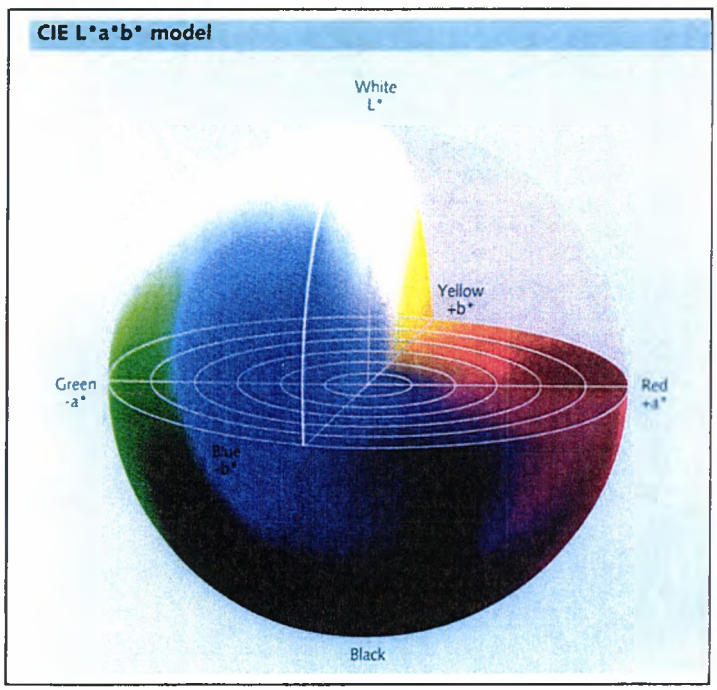


Figure 3.2 CIE L*a*b* model (Source: Agfa)

3.2.2. Colour Models used for Skin-Colour Detection

Observation of people's face suggest that skin colour can vary for a variety of reason and a range of colour-space models have been used to model chromacity and intensity to suit a number of conditions. Gong et al. (2000) note that intuitively large variations between people of different ethnic groups might be expected and an individual's facial colour might vary with temperature (e.g. when blushing). However, despite our own intuition, skin colour is closely clustered in 2D space such as HS (Hue-Saturation) and in fact most of the variation is in the intensity of the signal. Angelopoulos (2001) identified that there is uniqueness to the colour of human skin and measurements showed that that the overall percentage of light that was reflected from human skin increased with wavelength (Appendix II). The

University of Oulu Physics-Based Face database (2001) has results of skin spectral reflectance characteristics also had similar results. The haemoglobin level is the main cause of this special characteristic.

Early work on extracting facial images from complex backgrounds used texture models with monochromatic images (Dai and Nakano, 1995). When the technique was applied to RGB images just the I-component from RGB to YIQ conversion was used. Schiele and Waibel (1995) reported on using chromatic 'rgb' values so as to ignore the intensity component inherent in RGB values. Subsequently, colour was based on the two-dimensional representations of (r, g) the chromatic colours. Using these pair of chromatic colours a probability density function of the chromatic colour of an image was formed. Also the colour-maps of a range of people and skin types were obtained. It was observed that the face colours are located in a relatively small 'bandwidth' of the (r, g)-values. A general colour-map was produced that contained most face-colours. The pixel in an image with a given probability was compared with the general colour-map to give face-colour regions with higher or lower probability of being face colour. A binary image of the face area was found after averaging and thresholding.

Dai and Nakano (1996) isolated and retained the orange-like regions in the YIQ colour space as the skin region and eliminated the remaining regions. They then used texture in the grey image to identify faces in skin regions. The 'Pfinder' project (Wren et al, 1998) uses colour space on a blob, which is a cluster of spatial, and colour vectors. The use of YUV and normalised YU*V* components were used to remove shadows in a relatively static scene.

Chen and Chaing (1997) designed a face detection system based on skin colour classification by a neural network. Normalised values of x and y from the CIE-xyz co-ordinate system were used as input to a feed-forward three-layer network. The training method was based on a simple back-error-propagation learning method. The system was trained with fifty images. A pixel was tested to be skin-colour or not by inputting the 'xy' values to the input and the output thresholded to determine whether it was skin-coloured or not. Holes in the resulting skin-coloured region were caused by noise, highlights and other facial features (mouth, eyes etc.). Holes were filled by a range of filtering and segmentation steps to produce a 'candidate face region'. In order to detect the lip area more reliably the colour space was changed to L*a*b* colour coordinate system because of its uniformity in measuring colour differences. Experimental tests were conducted on three different groups of images. Images came from a digital camera, a scanner and from video or Internet sites. The first two groups gave 96% detection successes, but fell to 76% with the video images. It was concluded that the faces that could not be detected correctly usually had too many highlights or were too dark.

Cai and Goshtasby (1999) produced a method for detecting human faces in colour images using a chroma chart. CIE LAB colour space was used because it is perceptually more uniformly spaced than colours in RGB or HSV spaces. The CIE LUV colour-space model has the same characteristics and could also be used. Only the chroma, 'a' and 'b' components were used to separate skin from non-skin regions. Each pixel assigned a weight showing the likelihood of being skin. However, a training process is required to set up the system and used samples from

European, Asian and African people. The colour image is transformed into a grey scale image of probabilities of skin-colour.

Kjeldsen and Kender (1996) achieved hand segmentation based on a histogram-like structure called a Colour Predicate (CP). The predicate, once trained, was based on identifying candidate pixels with values indexed by HSI value. It was found that pixels with large or small intensities were discarded as hue and saturation became unstable in this range. Lighting effects produced changes in intensity, which produced changes in hue and saturation. A solution to this problem was to quantize the Hue and Saturation axis of the CP much finer than the intensity axis. Results indicated that segmentation worked well on a range of skin-tone but for optimum performance it should be trained on the people to be segmented. The technique was used on a range of video material, and worked well where there was constant unmodified lighting and did quite well on cluttered scenes such as crowd shots showing many people.

Chen et al. (1995) prepared a colour chart in HSV colour space that represented probable skin colours. Using 3 templates, they located faces in skin regions through a fuzzy pattern-matching algorithm. Sobottka and Pitas (1996) detected skin regions using Hue and Saturation and then selected regions that were elliptic as face regions.

A system for face recognition in dynamic scenes was described by McKenna et al. (1998). They also observed that human skin forms a relatively tight cluster in colour space even when different races are considered. Colour distribution in faces was shown in hue-saturation (H-S) space and modelled as Gaussian mixtures. Some problems were found by the large changes in the spectral composition of the scene illumination and it was found necessary to use at least use two colour models, one for interior lighting and one for exterior natural daylight.

Perez et al. (2002) made colour models by histogramming techniques in the HSV colour space in order to decouple chromatic information from shading effects. Colour information was found to be reliable only when both the saturation and value were not too small and set the thresholds at 0.1 and 0.2 respectively. Chen (2003) used a simple skin detection system of R>G>B, which could limit its use if white balance settings are incorrect. However, other colours, not associated with skin-colour, detected in this range were constrained by the use of an associated motion cue.

Sigal et al. (2004) introduces a novel approach to real-time skin segmentation in video sequences, despite wide variations in illumination. The system uses an explicit second order Markov model to predict evolution of the skin-colour distribution over time. Histograms are dynamically up-dated based on feedback from the current segmentation and predictions of the Markov model. The parameters of the discrete-time dynamic Markov model are estimated using Maximum Likelihood Estimation and also evolve over time. The performance of the algorithm for about seventy percent of the test sequences was much better than static segmentation. The algorithm is depended on the initialisation phase and is more susceptible to skin-colour background patches. Interestingly, the paper's review re-emphasises a number of points of previous work on skin-colour detection.

- Normalised RGB and HSV are the most common colour spaces used and are shown to be tolerant of minor variations of illuminant. These colour spaces tend to produce the minimum overlap between skin-colour and background- colour distributions.
- Parametric statistical approaches to skin-colour distribution, such as a gaussian model, have low space complexity and relatively small training sets. The major difficulty is order selection and is generally determined heuristically, and in constrained environments the model order can be predefined on the known environmental conditions.
- Although histograms that are used to represent density in colour space, and probability density can be evaluated easily, a major drawback is that a considerable amount of training data is required.

3.2.3. Colour Model Comparisons

Poynton (1997) explains that similar set of models, HSB (Hue, Saturation and Brightness) and HLS (Hue, Lightness and saturation), should now be abandoned. Nowadays when colours can be chosen visually, or related to other media, numerical methods like L*u*v* and L*a*b* should be used as they are perceptually based systems. Furthermore, Poynton points out that these types of colour space have a number of disadvantages i.e.

- The 'lightness' type of term makes no reference to the linearity or non-linearity of the underlying RGB and makes no reference to the 'lightness' perception of human vision.
- If the 'lightness' value is computed as (R+G+B)/3, it conflicts badly with the properties of colour vision, as it computes yellow to be about six times more intense than blue with the same 'lightness' values.
- These colour models are not useful for image computation because of the hue discontinuity at 360°.
- These models involve different computations around 60° segments of the hue circle and introduce discontinuities in colour space.
- Although these models appear to be 'device independent', the ubiquitous formulations are based on RGB components whose chromaticities and white point are undefined.

But not all researchers agree with Poynton's conclusions and there are instances of comparisons being made between different colour-space models. A valuable comparison of colour-space models was undertaken by Lee et al. (1996) for locating the human face. The experiments involved taking an ellipse of a typical face to analyse the results by various methods. Firstly a RGB scatter diagram was produced and showed that the data vectors are distributed along the diagonal axis and constitute one cluster, so the regions of mouth and eyes are indistinguishable. Three other colour space models were experimented with, HSI, L*u*v, and the principal component coordinates by Karhunen-Loeve transformation. The HSI colour space model showed clearly separated regions; the mouth was separated from other regions on the Hue axis and the skin, eye and eyebrow regions are clustered in the hue-intensity plain. The skin regions were concentrated in the low levels of the hue region, whereas the data for the eye and eyebrow regions were scattered in the mid

range of the hue regions. The CIE-L*u*v* coordinate system surprisingly gave clusters in the face region did not have a large colour distance between them. In addition using the principal component coordinates (KL space) from the eigenvalues and eigenvectors of the covariance matrix of the colour image values did not produce the large discrimination power expected. Further techniques for classifying the data from RGB, Lu*v* and KL spaces were investigated but still good classification results were not produced. The conclusion was that the HSI colour coordinate system was the best candidate for colour image segmentation.

Sigel et al. (2004) justify the use of the HSV colour model by citing the work of Terrillon and Akamatsu (1998). Terrillon and Akamatsu compared the performance of nine different colour spaces that found that the best were HSV and normalised RGB. Slightly worse discrimination was observed for other colour spaces. The disadvantage to HSV was noted as the costly conversion from standard RGB. However, in this work HSV was quantised into (64 x 64 x 64) RGB to HSV lookup table. It was noted that the paper referred to the HSV model in the text, but the figures referenced the HSI model. Some of Poynton's objections to the HS colour models have been addressed. Hanbury and Sera (2001) noted that the HLS colour space is widely used in image analysis, as it is physically intuitive. A new saturation-weighted hue order, which takes hue and saturation into account simultaneously, is discussed.

The discontinuity of Hue about its origin was addressed by Harding and Ellis (2003). Skin-colour in scenes that are properly white-balanced is usually in the red-orange region of hue. However, skin colour can appear different to this and often with a blue tint due to the illuminant or incorrectly set white-balance adjustments. Mathematical averaging does not give the correct result for a region that transverses this discontinuity. To alleviate this problem with skin-colour regions, the discontinuity was moved to the cyan region, and as a consequence the hue range was changed from 0 to 1 (or 0° to 360°) to +0.5 to –0.5 (or +180° to -180°), and discussed in the next section.

However, even with these disadvantages the HSV model has been found to out perform the apparently linear perceptually based CIELUV and CIELAB models as discussed in the previous section (Terrilon, 1999 and Lee, 1996).

3.3. The HSV colour-model

Smith's (1978) seminal paper considered two colour space models. One model was based on the hexacone (HSV) and the other based on a triangle model (HSI). Gonzalez and Wood (1992) later refined the latter model. The calculations for the HSI model and the algorithm for the HSV model are shown in Appendix II.

The HSV model avoids trigonometrical values and for 8-bit colour depth gives identical results of Hue as when the HSI model is used (Appendix II). In this model the RGB cube is projected along the grey vector onto a plane perpendicular to the vector, a hexagon disk results. The disk for black is just a point, but as the grey level changes toward white the hexagonal disk becomes larger. For each value of grey level there is an associated sub cube of the colour cube. The length of the side of the

colour cube in the projection is equal to the length of the side in the solid. V is specified to be equal to R, G or B and none is larger so that V = max(R,G,B).

It is instructive to compare hue values at the extreme of each sextant relate to the relative proportions of RGB values, as shown in figure 3.3. In the hexacone model, H and S specify a point in the disk for a particular value of V. H is taken to be the angle and S is taken to be the length of a vector centred at the grey point. The loci of constant S are hexacones, so when reference is made to the angle H, a proportional length along these loci is inferred. S is a relative length proportional to the longest possible radius at a given angle. S varies from 0 to 1, with the 0 value implying it to be a grey value and the 1 value implying it is a colour on the bounding hexagon. This also means that one of the values of R, G and B must be zero.

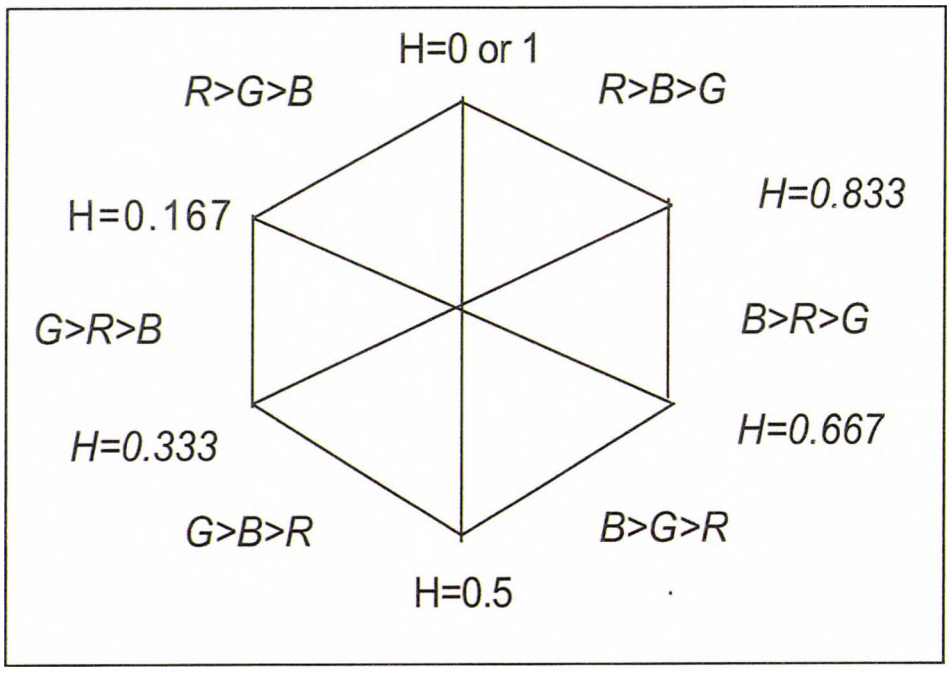


Figure 3.3 RGB proportions and related Hue values

The main problems when using the HSV colour model are singularity and discontinuity. Special care must be exercised at the singular point of Saturation, S=0 i.e., where R=G=B, the grey or achromatic axis of the hexacone. Hue is not defined along this axis and so cannot be used, as there is no colour and the reason of using the Hue is lost. In addition, if saturation is S=1 then one of the primaries must be zero. Experiments have found that any HSV processing should always run with some nominal default range of saturation so as to avoid these problems. A range of 0.05 to 0.95 has been found acceptable. At values outside this range, the calculation of hue can become unreliable due to the very small differences in the primary RGB components.

The discontinuity at the 0 – 1 boundary can give mathematical complications when calculating statistical values from samples that cross this region. This can be a complicated when the camera (video/web-cam) white balance has not been set properly. Automatic white balance mechanisms can be fooled when the background is predominately of a yellow/orange hue, like wood or a yellow wall. Resulting images have a blue tint and hues will spread from typical 0.05 values into the 0.9 to

1.0 range. A method of calculating average values of distributions that straddle the discontinuity is to move it to a different position. Moving the discontinuity to the cyan region, so the Hue range is now +0.5 to –0.5 corrects the difficulty for skin-colour detection. Figure 3.4 shows the new range of hue values and the formula to calculate colours in each sextant.

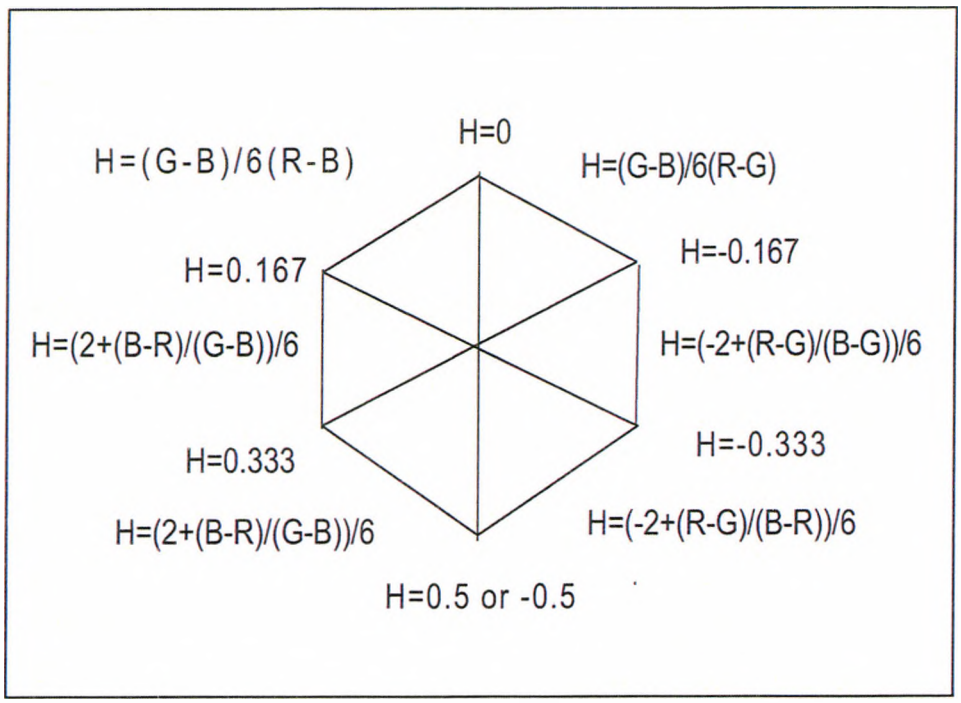


Figure 3.4 The hue hexacone using positive and negative numbers

3.4. Colour Space Experiments

Experiments with the HSV model (Harding, 1999) confirmed that the Hue of HSV remained very constant for a range of environmental conditions. Varying sizes of samples were taken from the back of the hand and also for a range of aperture settings of the camera, in fluorescent light. As would be expected as the aperture became smaller and less light was transmitted, R, G and B values became smaller. Converting the RGB values to HSV values showed that the V values and S values also decreased with a smaller aperture. However, the Hue value remained virtually constant, until the V value dropped to about 0.15 corresponding to RGB values around 40, where the H value increased slightly. In another experiment with tungsten light, when the light in some regions was intense at some wide apertures, upper values saturated the conversion to HSV values showed that Hue remained constant across a range of aperture settings. It was also confirmed that Hue values, under a particular illuminant stayed relatively constant when samples were taken from six men and six women. It was observed that the S and V values varied considerably form one person to another.

The previous experiment showed that Hue remained constant, for a range of light levels, but the mean value was different for the two illuminants. The fluorescent light gave a value of Hue at about 0.1 and the tungsten light at about 0.06. It should be noted that the white balance adjustment remained fixed in these two experiments and

was not necessarily balanced for the lighting conditions used in the experiment. It is observed that there is little in the literature that mentions the importance of white balance setting as this can affect the absolute value of hue recorded by the system.

Further experiments have been conducted to show the affect that illuminants have on the absolute value of hue. Appendix II shows a scene that is illuminated by four different illuminants. The illuminants were tungsten, white fluorescent, D50 and D65 lighting standards and gave average values of Hue, for a skin-like coloured region of the arm of a doll as, 0.0671, 0.1079, 0.1232 and 0.0767, showing the variation of Hue with changes in spectral light.

A number of experiments have been conducted of subjects in a range of environmental conditions as detailed in Appendix II. One experiment was based on the skin regions of an avatar. The result of sampling the 'skin' areas (hands and face) of the avatar gave values of Hue to be almost identical at 0.03, which is typical of that expected of human skin and in the region of R>G>B. However, another image, of the author, shows that sampling of the each hand and the forehead can give slightly different hue values for the three regions, but typically about 0.06, with the head region having the larger variation in hue value. It was found that virtually all skin-colour samples were within plus/minus two standard deviations of the mean, for a particular sample. This is consistent with a Gaussian model of skin-colour that 95% of all samples are within plus/minus two standard deviations of the mean value. If the range of Hue values were taken as the minimum of the three samples and the maximum of the three samples, then good segmentation of all the skin-coloured regions took place. A further refinement in segmentation was possible if the maximum and minimum of the Saturation values were included in the mask. The saturation variable helped isolate the skin-colour regions from the background but the mask area was generally smaller and typically excluded some eye and mouth regions.

Other image sequences were experimented with and either had poor white balance or were affected by lighting in some way e.g. a magnolia background affected by late afternoon sun. In these cases some of the hue values straddled the red discontinuity (red equal to 0 or 1 in unmodified HSV colour space model) and gave negative values of hue. Finally, a more complex image was obtained from the PETS database, in which there are three people. Taking samples from the arms and cheek of the person in the middle of the scene gave a range of hue values that allowed segmentation of all the skin areas in the image, except the right hand person's forehead. Sampling of the forehead showed that hue values were in the negative sextant where R>B>G instead of the more normal situation of R>G>B. Interestingly, this person's forehead appears different to other skin tones in the image. Considering the context of this experimental sequence, then the person could be experiencing 'stage fright' and the change of skin colour is often observed. It interestingly links to the earlier discussion about the unique spectral characteristic of human skin. If the 'fright or flight' syndrome is pumping more blood around the body it is likely to affect the haemoglobin that affects the skin's spectral response.

3.5. Fusing of Motion and Skin Colour Cues

Motion can be detected by taking the difference between the foreground or object and the background. Many motion detection schemes incorporate some background updating approach to avoid errors in segmentation due to variations in lighting or the background. Stauffer et al. (1999) overcome this problem by modelling each pixel in an image as a Gaussian distribution. The program usually has the ability to model a number of distributions, depending on the expected variability in the background which can be quite severe in outdoor situations. The Gaussian distributions are compared with each other to determine which may correspond to background colours based on their persistence and variance. The Gaussian distributions can be modelled in RGB, rgb (chromatic) or HSV, space etc. The pixel values that are not matched to a distribution are considered part of the foreground. Eventually these pixels will be absorbed as a Gaussian distribution if there is sufficient evidence supporting them. It is common to have an updating rate or learning rate that can vary the frequency of updating the weights in the models. The equation controlling the updating acts like a causal low-pass filter with an exponential window on past values.

Vermaak et al. (2003) tackles the importance of incorporating adaptivity to observational models to counter, for example, illumination changes that affect surface colour. The rate of adaptability, that is applied, is set so as not to be affected by transients. Adaptation is only allowed during the conditions of the object being present and in motion. This particular method uses combined colour and motion observations from a fixed filter bank, with motion used to initialise a Monto Carlo proposal distribution. Adaptation is performed using a stochastic EM algorithm during the periods detailed above.

KaewTraKulPong and Bowden (2000) address the problem of using appearance and motion models in classifying and tracking objects when detailed information of the object's appearance is not available. Objects are associated temporally by using motion, shape cues and colour information. It is explained that when the number of pixels supporting an object is too few to train a complex shape or colour model, it is found unreliable to learn just a colour distribution due to the limited number of examples from the scene. In addition if the model becomes too complex, the number of training samples increases exponentially and it is unreliable to classify or track objects by shape or colour alone. In the example of tracking of walking people with low-resolution images the key strength of the technique was in the use of robust background modelling and colour mapping obtained from anthropological study to model low resolution colour targets. The motion and colour information was combined using probabilistic methods and the system was able to track multiple people moving independently and was able to recover from lost tracks due to occlusion and background clutter.

Sherrah and Gong (2000) demonstrate that robust solutions to computer vision problems can be provided by perceptual fusion. This can be achieved through a framework for fusing different information sources through estimation of covariance from observations and is demonstrated in a face and 3D pose tracking system. Siebel and Maybank (2002) explain that tracking can be classified into three main categories of increasing complexity i.e. region or blob-based tracking with additional classification systems based on colour, texture or other local properties; 2D

appearance based models and 3D models. In their description of a people tracking system, four co-operative modules are used i.e. motion detection, region tracking, head detection and an active shape tracker. It was found that by fusing the output of each output a high tracking reliability can be obtained than any of the individual trackers can achieve on its own.

In the gesture experiments an adaptive background technique similar to Stauffer and Grimson (1999) was investigated but found unnecessary for most gesture sequences. Firstly gesture sequences were recorded indoors where there was little change in the background. Secondly the gestures were quite short, lasting just a few seconds, so updating could significantly lag behind the foreground motion and affect the output giving inaccurate motion segmentation. In order to have an output that located motion, traditional motion detection could be used by detecting the difference in intensity in two adjacent frames (Jain et al, 1995) and then forming a binary mask image by selecting a suitable threshold. Two possible masks are possible, the DP (Difference Picture) or the ADP (Absolute Difference Picture) as defined in the following equations: -

$$DP_{jk}(x,y) = 1 \quad \text{if} \quad F(x,y,j)-F(x,y,k) > \tau$$

$$= 0 \text{ otherwise}$$

$$ADP_{jk}(x,y) = 1 \quad \text{if} \quad |F(x,y,j)-F(x,y,k)| > \tau$$

$$= 0 \text{ otherwise}$$

where τ is a threshold, and j and k are two images at different times.

Fusing colour and motion together has many advantages as it reduces the variability in each of the individual cues. As already discussed, McKenna et al. (1998) found that colour models had to be changed when the scene illuminant changed. The following technique has a high tolerance to colour change especially when adding the motion cue to the colour cue. Additionally, only a limited amount of thresholding is used and a novel rank-ordering of skin-colour and motion objects is employed to successfully work without adjustment in a range of sequences and illuminations.

The fusing of colour and motion cues is reported by Chen et al. (2003) and Harding (2003). They showed that motion objects and the skin-colour objects, as a result of hand movement, overlap. The hue (or hue and saturation) mask of skin-coloured regions, HS is obtained by: -

$$HS_t(x,y,t) = 1, if, h_t > H1, and, h_t < H2$$
$$HS_t(x,y,t) = 0, otherwise$$

where 'h_t' is the hue (or hue and saturation) of the current frame and H1 and H2 are the hue threshold values used for segmentation, as discussed previously. If saturation is also incorporated in the segmentation then two other threshold values S1 and S2 are also required. The application of a logical AND function to the motion object mask and the skin colour masks to form the skin colour and motion mask and related objects (SCM), henceforth referred to as gesture objects: -

$$SCM(x,y,t) = ADP_{t,t-1}(x,y,t) \,\&\, HS_t(x,y,t)$$

The reasoning behind this method is best explained by considering the movement of an object and in this context would typically refer to the hand movement. To explain the concept, the object is considered non-deformable and the intensity change between the object and the background is significantly different. Figure 3.5 shows the intensity profile of the moving object in frames one and two, F(x,y,1) and F(x,y,2) respectively and shown as binary valued for the sake of simplicity. The absolute difference picture is shown as ADP_{12} , which is also shown as binary valued as a result of appropriate thresholding. The Hue-Saturation mask, HS_2 of the object at frame 2, is shown in this example to have the same profile as the grey-scale intensity profile, which is also shown as a binary profile as a result of thresholding. The motion cue and skin colour cues are simultaneously shown to occur by the logical AND of the ADP_{12} binary profile and the HS_2 binary profile and are referred to as SCM objects. It is observed that if there is little movement the duration of the gesture object is relatively short compared to large movements. However, whatever the duration of the gesture object the implication is that the object is a consequence of motion of the skin-coloured region.

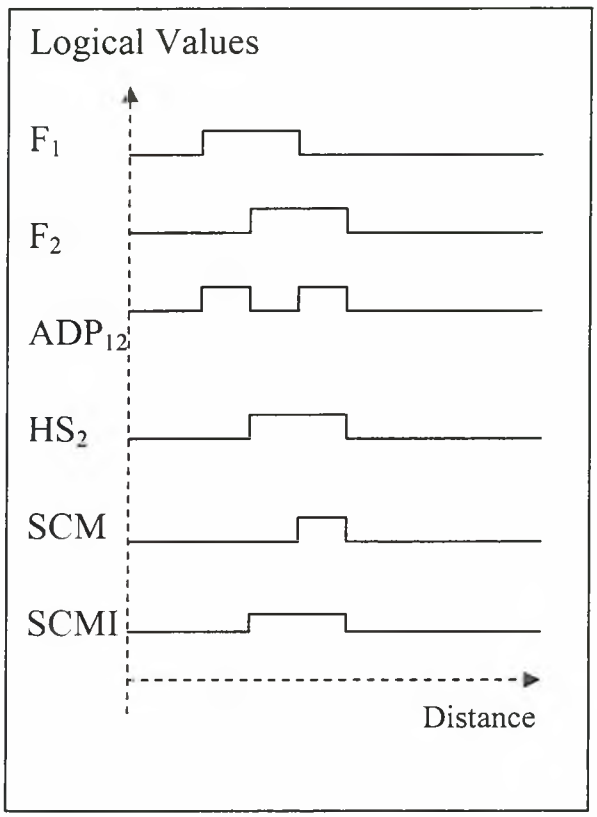


Figure 3.5 The cross-sectional profile of a simple moving object is shown at frames F1 and F2. The thresholded absolute difference picture 'ADP_{12}' is logically ANDed with the Hue-Saturation mask HS_2 , obtained from F_2 to produce the skin-colour and motion mask, SCM. The SCMI mask indicates the original HS mask region congruent with the SCM mask.

As the gesture object occurrence is congruent with the skin-colour occurrence, as indicated by the HS_2 profile, it can be implied that the gesture object is a sub-set of the hand region. Hence the output labelled SCMI is a region/object identified by the gesture object located by the HS profile of the hand. In this idealised example there is just the one HS profile and the HS profile and the SCM profile appear the same. However, in a real situation there is the likelihood of there being more than one skin-coloured region and it is the gesture object that locates the skin-coloured region that is in motion. The performance of this technique is assessed with real images in the next section.

Chen et al. (2003) also devised a method of separating the arm region from the hand region, which can be a problem if the gesturer is wearing a short-sleeved shirt or blouse. In order to track the hand position, rather than some position on the arm, Chen et al. (2003) used the properties of a simple edge detector, the Kirsch operator to obtain different edge directions and then choose the absolute maximum value of each pixel to form an edge image. In general the edges of the arm are less than the edges around the fingers. The addition of another cue based on the amount of edge information logical was used to combine the skin-colour, motion and edge masks together to produce a mask with SCME objects.

3.5.1. Motion Experiments

It is pertinent to note that there are two basic objects formed due to the difference of the right hand in frame one and the background, and the difference between the background and the hand in frame two. Because the background can vary, the differences representing the two objects will be different, and hence after thresholding the areas of each object may be different. Furthermore, as the hands are deformable objects, differences in hand orientation, between frames, may cause differences in object size, after thresholding.

Colouring of the grey-scale images helps show where the maximum differences are realised for Difference Pictures and Absolute Difference Pictures, as shown in Figure 3.6, before thresholding takes place.

For each image sequence the threshold value needs to be set for optimum segmentation. Two threshold values of 0.05 and 0.5 are shown being used on the absolute difference picture, in Figure 3.7. The former is set at a value just above zero and the second is set based on the difference images of Figure 3.6. It is not always possible to have a manual system to determine optimum threshold value for a particular sequence of images. The next section introduces a novel technique to overcome this constraint and does not require the threshold value to adapt to different image sequences or conditions and can be left set at a minimum value, just above noise levels, for all image sequences.

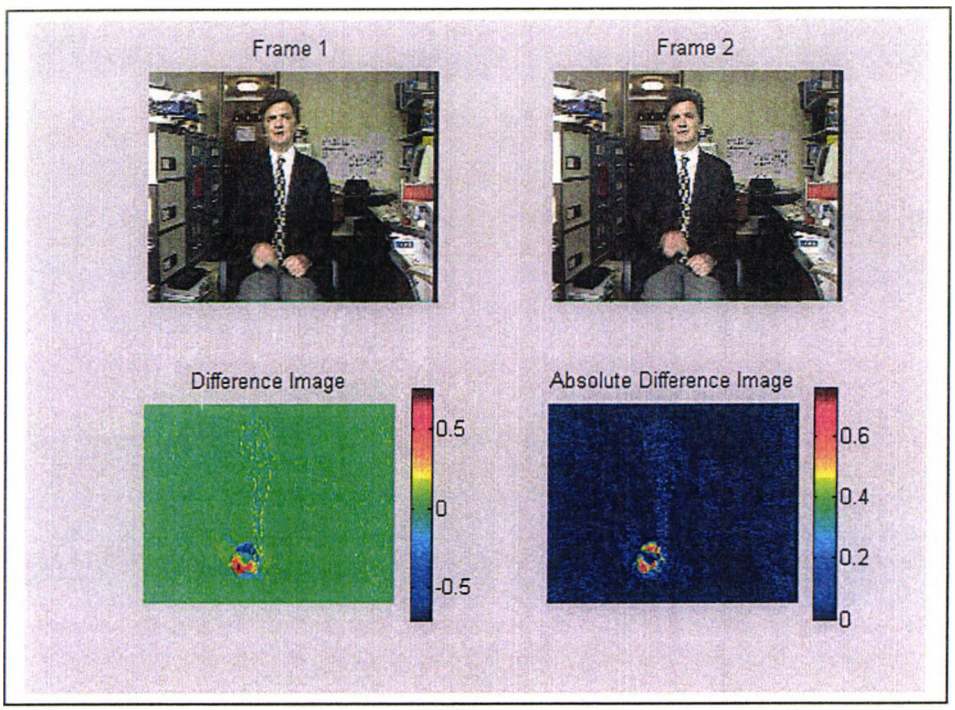


Figure 3.6 Difference and absolute difference pictures of frames 1 and 2 before thresholding

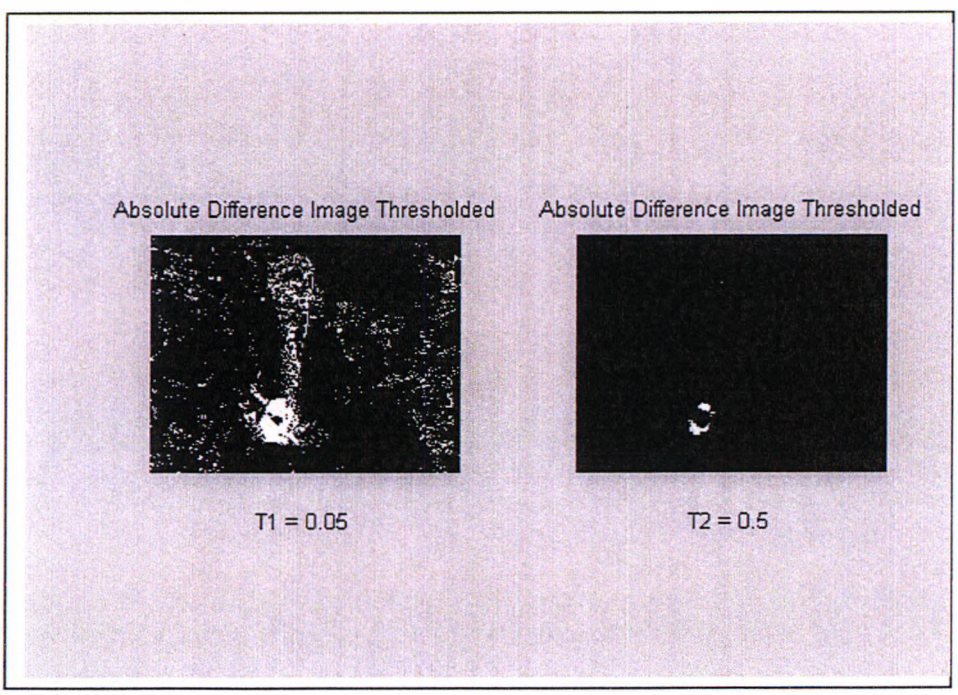


Figure: 3.7 Two different thresholds of absolute different pictures

3.5.2. Rank Ordering of Motion Objects

With a minimal threshold many objects are generated which are not related to the main gesture movement and represent small movements or noise. A size filter can

remove these small objects, but setting the threshold level for the minimum object size is an arbitrary or heuristic process.

Sorting the objects, from the thresholded, absolute difference picture, by descending area is a less arbitrary process. The largest object is the first object of the sorted data and usually relates to the largest motion between the two frames. Instead of having to decide upon a threshold value, a decision as to the maximum number of objects to process has to be decided. This is a far less critical decision to take and some six objects have been found to easily accommodate a range of situations.

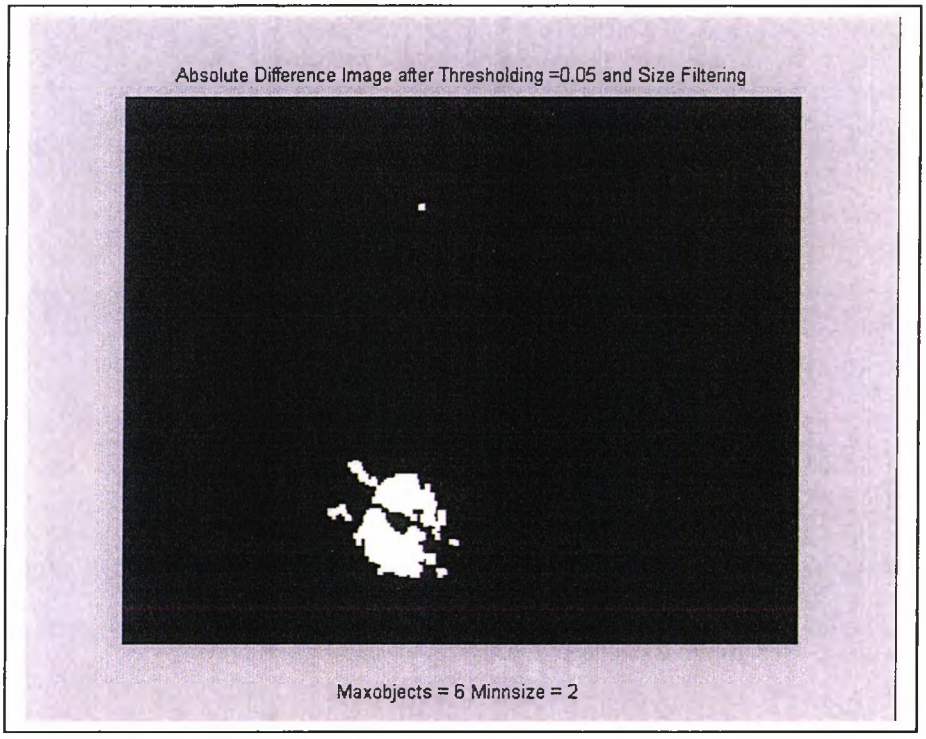


Figure 3.8 Six largest objects as a result of rank order by area filtering

The example shown in Figure 3.8 shows just the first six most significant objects. Coincidently the result is very similar to the thresholding at 0.5, as shown in figure 3.7. If required small objects can be removed by using a combination of minimum area criteria or morphological filtering such as 'opening'.

Rank ordering preserves the most significant object criteria regardless of image size. If the sequence was to be analysed at two different image sizes, the most significant object will be the largest object in each image although the actual area will be different. Again this method avoids the need to adjust size filter criteria depending on image size.

The data produced from the rank order by size filtering operation is shown in Table 3.1. The table shows data sorted in rank order with the original object number given to the raw binary data in the labelling routine, the area and the row and column coordinates respectively.

Object Number	Area	Row Coordinates	Column Coordinates
5	899	236	144
4	750	210	150
1	71	219	116
12	28	250	170
14	17	234	177
8	15	59	159

Table 3.1 Rank ordered size filtered motion data

The first two large area objects (Object No. 5 and 4) represent the movement of the right hand. The sixth number (Object No. 8) represents a small movement of the head. The other three objects of relatively small areas are due to motion associated with the hand movement. These motions can be excluded by the combination of a colour component as will be discussed in the next section.

3.5.3. Production of Skin-Colour and Motion Objects

Gesture objects are produced by the logical AND of the motion mask and the skin-colour mask. Figure 3.9 shows the Hue mask used with the motion mask of Figure 3.8 to produce the right-hand image of Figure 3.9. The number of gesture objects is generally much less (6), as shown in Table 3.2, than from the motion mask (at least 14). Furthermore, sorting the gesture objects by descending rank order of area highlights the most significant objects relating to the movement of hand.

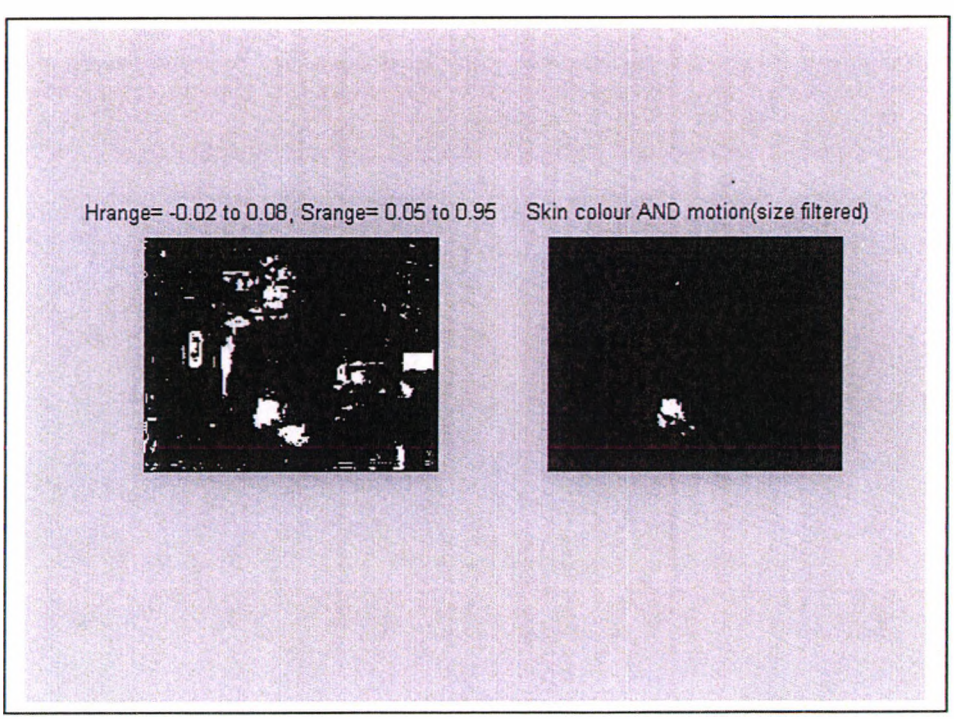


Figure 3.9 Hue-Saturation mask (left), gesture objects from SCM mask (right)

Object Number	Area	Row Coordinates	Column Coordinates
2	624	212	153
1	181	227	142
6	17	234	177
3	15	59	159
4	13	243	166
5	2	250	173

Table 3.2 Rank ordered size filtered gesture objects

3.6. Hue and Motion Segmentation Discussions

When the hand is in its most dynamic state, the distance that it moves between frames is considerable (at least a hand width) when images are captured at 25 frames/sec. In this dynamic state the gesture object is clear to see and is the most significant object in the image. However, when the hand is moving more slowly, typically at the beginning, end or at the top of its trajectory, other segmentation issues occur. Firstly, segmentation of the hand is not so complete and segmentation fragments from a single object to multiple, smaller objects. Secondly, when the dominant hand is nearly stationary, other movements of skin-coloured regions are just as likely to be detected. These skin-coloured are usually the other hand or the face. In some cases it can be the result of the exposure of a skin-coloured area in the background, that has become visible by movement of say the arm, but are usually of a transient nature.

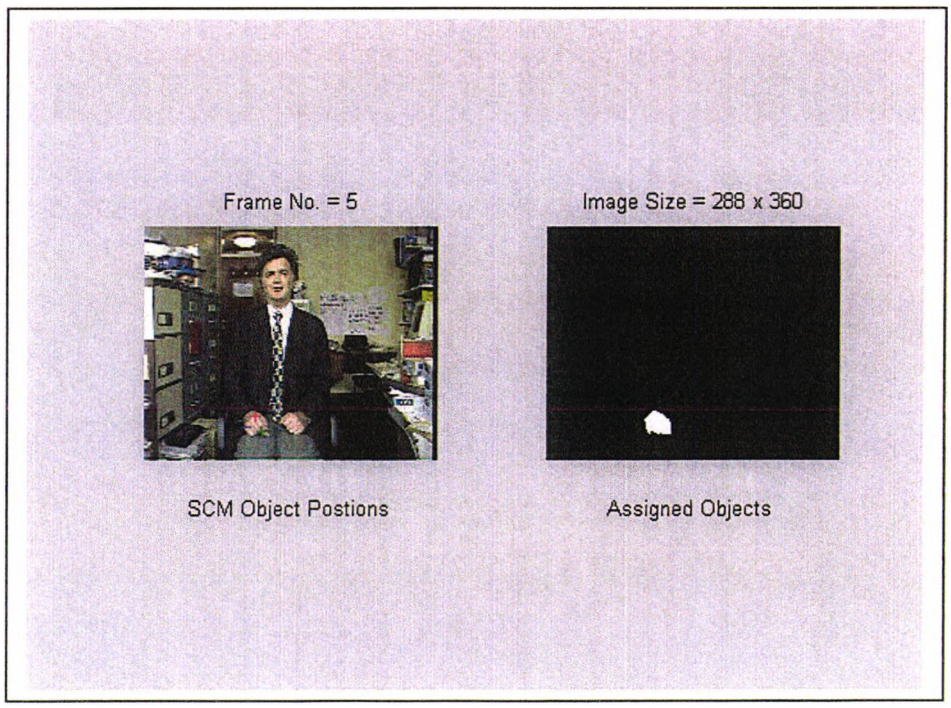


Figure 3.10 The position of the first three most significant objects shown by the red, green and blue crosses, respectively on the left image and the assigned or SCMI object in the right image (image size 288 x 360).

Fragmentation of the gesture object can give rise to the issue whether the most significant objects centre of gravity is the most important point to track. The most significant object is now more likely to be seen near the edges of the hand (fingertips and cuff), as this is where the most prominent change of intensity occurs. A procedure can be implemented where these objects are tested to see if they are located in the hand region given by the Hue mask. If they are, they are assigned to that region and produce SCMI objects. This results in fewer objects in the output and the ability to track the centre of gravity of the hand outline. Figure 3.10 shows the result of this operation.

Smaller images (144 x 180 instead of 280 x 360) do not change the situation, as shown in Figure 3.11, except the positions of the first and second gesture objects are interchanged, in this case. However, consideration of the size of the first object shows that it is some five times larger than the second object and closer to the centre of gravity of the assigned SCMI object.

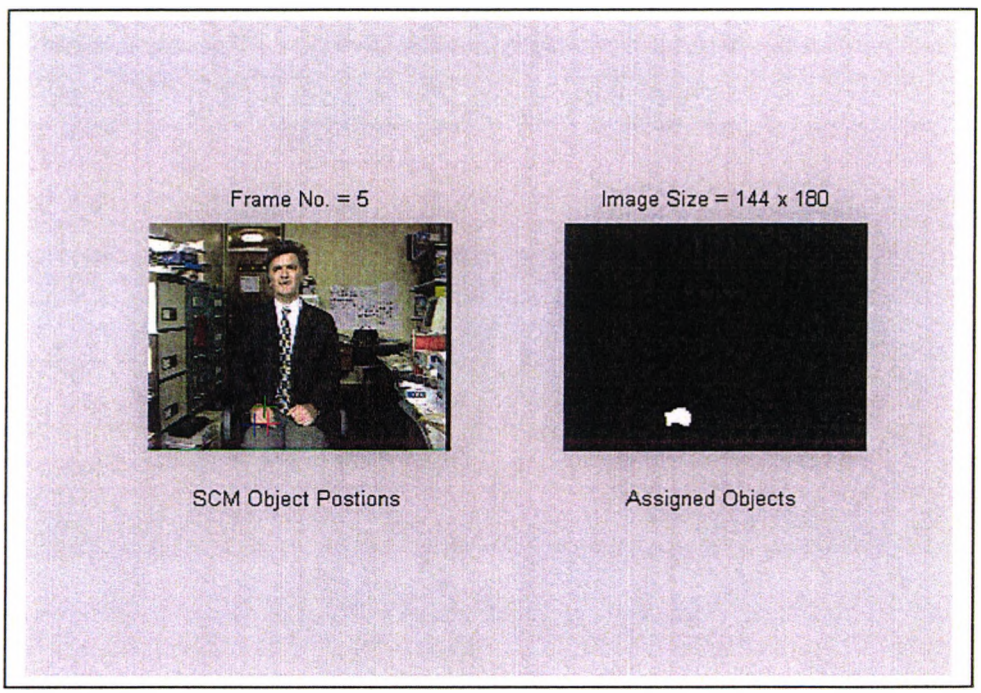


Figure 3.11 The position of the first three most significant objects shown by the red, green and blue crosses, respectively on the left image and the assigned or SCMI object in the right image (image size 144 x 180).

The main problem with the generation of SCMI objects is that the segmentation, by hue (and saturation), of the hand is critical. It is important that the hand is clearly segmented from the background, which is not always the case. It has been found that the hue range was not critical for hand gesture tracking. In figure 3.9, the hue range was set quite arbitrarily wide (-0.2 to +0.8), whereas in figures 3.10 and figures 3.11 the skin-colour was detected by using hue and saturation values (Hue range 0.053 to 0.094 and Saturation range 0.225 to 0.375). Although, a method of finding the Hue range has been discussed, the inclusion of the motion cue can make the hue range less critical.

In general it has been found useful to perform morphological 'opening' on the motion mask to remove noise or small apparent motions, which have little to do with

skin-coloured regions. However, it has been found advantageous not to filter the gesture objects because the smallest object can be relevant to detecting the smallest skin-coloured motion. At this stage, this information is not discarded but left to the object selection algorithm (Chapter 4) to decide if the information is important or not.

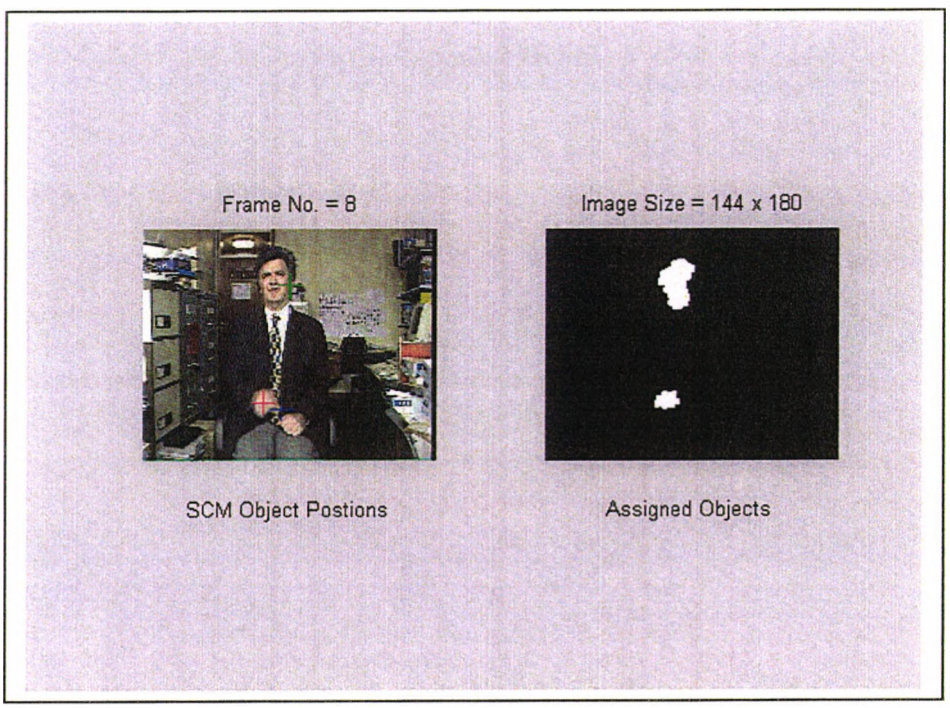


Figure 3.12 The second gesture object (green cross) in the left image detecting motion in the face region and shown as an SCMI object in the right image

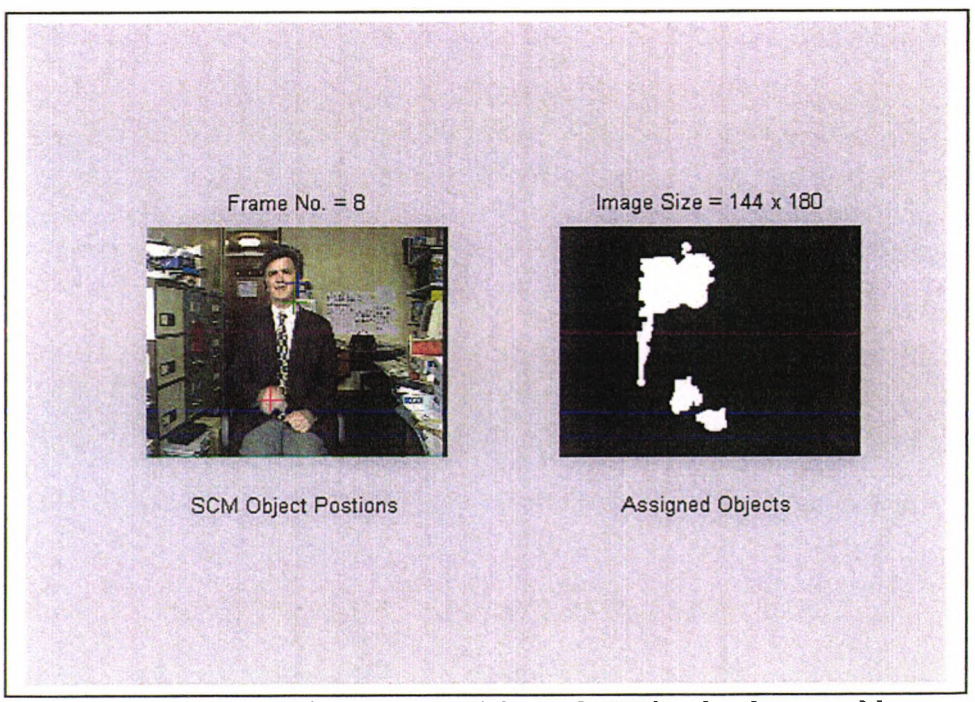


Figure 3.13 Second and third gesture objects detecting background because of poor segmentation.

The left hand image of Figure 3.12 shows the position of the first three gesture objects, although the second and third gesture objects are extremely small in size compared with the first object. In some circumstances such small objects could be ignored as small motions of the face may not be significant. However, the generation of the SCMI objects shows the face region and dominant hand region highlighted as shown in the right image of Figure 3.12. This segmentation could be important if hand silhouettes are to be used for recognising static hand gestures or signs, but not discussed in this thesis.

When the saturation range is relaxed to the default range (0.05 to 0.95), a number range of other objects are included in the face mask, as shown in Figure 3.13. Figure 3.15 shows that the skin-like colour of the wooden door has now been included in the segmentation. Interestingly, the assigned or SCMI objects show the left hand being also highlighted being due to the fourth very small gesture object, not shown in the left-hand image.

The issue to be resolved is whether it is important to track the centre of gravity of the overall shape of the hand, or any other part of the hand. This is best left to the next chapter when the performance of the detection system is assessed on whole sequences. There is evidence at this stage that it may not be important, as in the literature there has already been a comment (Howell et al., 2003) that too precise data is not realistic. Furthermore, the actual location on the hand may not be important as the change in position on the hand may be considered as a high frequency signal which can be ignored if the system is interested in mainly low frequency harmonics.

The advantage of fusing skin-colour and motion cues can be seen in Figures 3.14 and 3.15. The hue image is segmented into a range of 'narrow-band' overlapping hues ranges, as shown in Figure 3.14. In this particular example, the images were segmented in 'narrowband' hue range of 0.02, starting at –0.01 (magenta-red) and finishing at 0.1 (red-orange), and highlighted red in each image. The range typifies the expected skin-colour affected by different illuminants or poor white balance. The twelve different images of Figure 3.14 show how different colours in the scene are highlighted as the scan moves across the range of hues varying from orange to red to magenta. Combining the overlapping, hue ranges with the motion mask, produces the results shown in Figure 3.15, with a clear suppression of the reddish regions which highlight where a small movement of the hand has occurred.

Harding & Ellis (2003) suggested that an automatic calibration procedure could be implemented from this result. The largest region could be obtained from the images shown in Figure 3.15 and an updated reading of the colour of this region obtained. This attractive technique works well if the sequence is well lit and the only appreciable motion is from the dominant hand. However, the colour segmentation of twelve images is very time consuming and can only be practically be used in an initial calibration period.

Figure 3.14 Overlapping hue ranges of 'width' of 0.02 and starting at –0.01 (top left image) and finishing at +0.1 (bottom right image), incremented by 0.01 and shown in red.

Figure 3.15 Combining the overlapping hue ranges (Fig 3.16) with the motion mask to produce skin-colour and motion mask in red.

3.6.1. Combination of Colour, Motion and Edge information

Investigations as to the characteristics of the Kirsch templates revealed that an alternative set of edge templates were available, attributable to Prewitt. Experimentation as to possible threshold values was undertaken on an avatar sequence of images so that background effects were minimal and edge effects were clearly associated with the hand region. Results of the investigation resulted in a preference for the Prewitt templates, as the output was fairly insensitive to a range of threshold values. In addition, Kirsch templates required much higher threshold values outside the normal 0 to 1 range, and made it less suitable for a heuristic approach to segmentation. The logical AND of the skin-colour, motion and edge masks produced SCME objects and the SCMEI objects indicated by the SCME objects congruent with the skin-colour regions. However, this technique was only effective in well-lit image sequences as is discussed in the next chapter.

3.7. Hue and Saturation in other environments

Previous sections have concentrated on the segmentation issues in relatively well lit environment and images produced under appropriate white balance control. The skin-colour was typically captured by an heuristically set range of -0.02 to 0.08 and saturation values set at default range of 0.05 to 0.95. Experiments showed that skin-colour could be segmented using a hue range based on based on plus/minus two standard deviation from the mean, consistent with a Gaussian distribution.

The skin-colour was segmented in a range of other environmental conditions and is recorded in Appendix II. Further work on the sequence discussed in this chapter showed that as expected both hands gave similar average hue values of about 0.065, but the forehead was noticeably higher at 0.083. It was established that all skin-coloured regions could be segmented by taking the highest and lowest of the two standard deviations from the mean of the three regions. However, comparison with the colouring of the skin regions of an avatar gave interesting results. In this case the average hue values were all the same, averaging about 0.03, not that dissimilar to the previous results. However, the forehead and hands had exactly the same colour rendering. Another sequence, made available by PETS, gave hue values in the 0.01 to 0.047 range and again very similar to the sequences previously discussed. However, one of the three subject's forehead was not segmented by this method. Closer investigation showed a forehead very flushed and obviously redder that the other subjects. In this case the forehead skin-colour sample confirmed that the hue had moved into the -0.05 to 0.03 range.

Two other sequences were experimented with which generally had poor white-balance adjustments and poor lighting conditions of some kind. The sequence of a subject, in relatively low light conditions, gave skin colour regions that had shifted to redder colours with a range of hues approximately -0.01 to 0.04. Again natural wooden doors in the background could only be segmented out of the image by the use of saturation values. The other challenging environmental condition was illuminated by late afternoon sunset where the hue values of the skin-coloured regions shifted further into the negative values and the range of hues was typically -0.14 to 0.03.

3.8. Summary

There are many different colour-space models that can be used for detecting skin-colour. Through the auspices of the CIE, models have been developed that are similar to human perception. The most favoured model is the L*a*b model, which has a linear relationship with colour distances. It has been reported by two authors, that some perceptually linear models do not give the expected results for segmentation. Although there is some opposition to the HS variety of colour space models it seems to functions adequately for many computer vision applications.

Some authors segment the skin by the condition R>G>B, thus highlighting typical skin-coloured regions and rejecting five-sixths of all other colours, but does require a good white balanced image sequence. The benefits are in the processing speed as the conversion from RGB to HSV can be costly. However, the main advantage of the HSV model is that in representing a colour by Hue, a single variable that is easy to control and adjust. If an 'HS' pair or a 'rg' pair is used, then two variables have to be adjusted. Although it has been shown that an HS pair can segment a skin region from other similarly coloured regions, there is evidence that in a sequence of images saturation can vary from frame to frame more than with hue and hence cause segmentation to fail.

It has been found that the Hue of skin-colour stays very constant for a range of lighting intensities and remains very similar for a range of people. It is also shown that the Hue varies for a range of different illuminants. The physics of skin absorption and reflection shows that it has a unique spectral signature due to how haemoglobin absorbs light. In general the hue expected of skin, in a scene, properly white balanced, would be expected to be in the region of about 0.01 to 0.05 equivalent to hues being described as being biased toward the red end of a red-orange colour. The saturation range would typically be 0.25 to 0.35 indicating a considerable amount of white added to the hue. It was also found that the spread of skin-colour was maintained within plus/minus two standard deviations of the average value.

In a sequence of images provided by PETS, one of the three people in the image does not segment as well as the other two people for a given set of conditions. The forehead of this person appears redder than the other two, perhaps through embarrassment at acting in the scene. In these cases the hue of the forehead was found to shift toward red so the forehead of one person was not segmented without adjusting the range. A similar but less pronounced difference between hand colour and forehead colour was found from the experimentation of the sequence produced by the author. Interestingly, the image sequence of the author, discussed in this chapter, showed the forehead actually being more orange than the hand regions.

Investigations of the three skin-coloured regions of an avatar gave similar hue values, averaging about 0.03 which is about the expected value found from the two previous experiments, with saturation values just a little less than found experimentally. However, it was noted that to make the avatar more lifelike then the facial region should exhibit some different skin-colouring to that of the hands. The two experiments with poor white balancing both showed average skin-coloured hues that had shifted into the red-magenta regions and saturation range being substantially of the same order as before but perhaps some wider distribution of values.

In the experiments two important problems with the HSV model are observed. First is the discontinuity of the circular model of Hue where red can be dual valued at 0 or 1. Any mathematical averaging of samples taken in this region can give false values of Hue. The author has found that moving the discontinuity to the cyan region, and changing the hue range from 0 to 1 to –0.5 to +0.5 overcomes the mathematics problem. In addition, the white balance in images is not always perfect and skin colour can be detected in the negative sextant when R>B>G as opposed to R>G>B. The example cited of the person with a redder forehead has more blue content than the other two persons, and the hue range for segmentation has negative values. The second problem with the HSV model is singularity. When saturation is zero all primary colours are equal value and the result is a grey or achromatic condition. Hue is immaterial at this point, but small deviations of the primary colours can result in quite bizarre colours being produced. Similarly, when saturation is at one, one of the primary colours is zero. These effects can be reduced by using default values of saturation that removes the influence of small difference in the primary colours.

Many of the skin colour techniques rely on comparing pixel values with probability density functions of typical skin-colour pixels, which have been obtained through training from experiments or from the Internet. It is often cited that training time can be considerable. This paper explains a system that does not require training, at the worst some initial calibration of hue either manually, or a semi-automatic procedure can find a suitable hue range for segmentation. However, it has been found that this is not critical (rather like the condition of R>G>B) and an educated or experienced guess at hue range can work well, as the motion cue greatly reduces the amount of possible objects in the sequence that are skin-coloured and moving. Further refinements to this technique can be achieved in some sequences by using edge templates to help locate the finger region. This type of segmentation does not always produce an object that represents the complete hand, but typically multiple small objects are produced that would lie in the hand region. An enhancement to tracking the hand is to indicate where these SCM objects originate from in the HS mask and form SCMI objects. This can reduce the number of gesture objects and allows the centre of gravity of the hand to be tracked more accurately if lighting conditions and associated skin-colour segmentation is good.

The tracking of the rank ordered skin-coloured, moving objects is considered in the next chapter. In chapter four, sequences recorded under different lighting and environmental conditions are analysed to show how well the technique works for one-handed gesture sequences and how it could be extended to two handed gesture sequences. It is shown how some simple object selection algorithms can improve the quality of data and prepares it for time domain normalisation prior to obtaining the frequency content of the gesture trajectories. The data is also compared with ground-truth data to give a good comparison of the technique and indicate its accuracy.

4. Time Domain Tracking and Normalisation

The investigation of the ability to track hand trajectories in a complete gesture sequence by gesture objects links this chapter to chapter 3. A number of changes of different parameters is experimented with and conclusions formed. It was found that environmental lighting conditions affected considerably the production of gesture objects. In some conditions the second and third most significant objects needed to be considered in the tracking process. In order to differentiate as to which object to track from frame to frame an 'Object Selection Algorithm' was developed. A consequence of this work was that it was found possible to independently track both moving hands and also to track a hand moving in scenes in which there were other significant movements by other people. The resultant tracking coordinate output of the gesture was of variable length due to the variability in gesture and gesturer. To enable frequency components to be effectively compared each gesture needed to be normalised to a set length. This normalisation was achieved by using multi-rate methods of interpolation and decimation.

4.1. Introduction

Following on from the previous chapter, which established the necessity of locating the hand from cues generated by colour and motion, this chapter examines how these cues perform in a sequence of images that comprise a gesture trajectory. There are three possible outputs to appraise.

The first is to judge the effectiveness of the fusion of the skin-colour and motion spatial coincidences, to produce gesture objects (skin-colour and motion). The gesture objects are produced from a mask that has been generated from the logical AND of two other masks. These two masks are a result of applying suitable thresholds to difference images to capture motion and Hue-Saturation images to capture skin-colour. The second approach is to take gesture objects (SCMI) produced when SCM objects are coincident to the objects produced by the segmented skin-colour image, which normally would be expected to show objects related to the hands and the face regions. Finally, the finger region is detected by an edge template mask and combined with the two former masks of skin-colour and motion to produce SCME or SCMEI objects.

During these experiments, it was appropriate to examine the role of the Hue and Saturation segmentation range for a sequence of images. The Hue range for detecting skin-colour, in many environmental conditions, may be non-critical, and a Hue range based on experience, can save calibration or training time. Basically, a Hue range that reduces the total colour range reduces the number of coloured, moving objects that can be encountered. The use of Saturation outside a nominal range can isolate particular colours quite successfully, but is prone to change if for instance lighting changes or lighting is uneven across an image. This is usually less of an issue for short gesture sequences, but can cause segmentation of skin-colour to fail more frequently than if just Hue is used.

The set of rank-ordered objects does not always produce the most significant object related to the dominant gesturing hand. First, there is the occasion in a gesture trajectory when no motion is detected. No motion occurs when the hand is stationary,

typically at the beginning and end of the trajectory and at the peak of the trajectory or gesture. Additionally, the camera/video mechanism may not capture a difference between frames, and the movement of the hand toward the camera may appear stationary in the appearance–based view. Additionally, as the gesture tends toward a static or partially static condition, the gesture objects tend to fragment into a number of smaller objects because of less distinct movement. In these conditions the other skin-coloured objects (usually the less-dominant hand or face) can produce a gesture object that is more significant in the ranking order than the dominant hand. When these conditions occur a decision needs to be taken to decide which object to follow or assign to the dominant hand. In ideal conditions, with no other movement coming from the gesturer and no noise objects being generated, the gesturing hand location could be determined by the coordinates of the most significant object, which may also be the only object generated.

However, with less than ideal conditions there are several objects generated and the most significant object may not always relate to the dominant hand. The Object Selection Algorithm (OSA) was developed to determine which object to follow to track the trajectory of the dominant hand. The success of the OSA is judged against 'ground-truth' data that was derived from image sequences by visually determining the hand position and recording the coordinates of the approximate centre of gravity of the hand's coordinates.

The output of the OSA then needs to be formatted for input to the FFT to generate harmonic frequency components. This is explained, in detail, in the next chapter. Before this process can take place, the data sequence needs to be normalised which is a technique akin to 'dynamic time warping'. However, the process must not only resample the spatial data it must be a linear process to preserve the time element. Normalisation is necessary, as harmonic components can only be compared from sequences of the same number of samples. As gestures can vary in duration, by the gesturer and from gesturer to gesturer and by different frame rates, this normalisation is imperative for analysis to take place. Normalisation is achieved by multi-rate methods that adjust the number of samples in a sequence by interpolation and decimation techniques.

4.2. Previous approaches to tracking

The tracking of objects over a series of frames is relatively easy if detection is robust and there is no occlusion. In the case of many entities moving independently, tracking requires the use of constraints based on the nature of objects and their motion (Jain, 1995). First, it must be assumed that sufficient samples are taken of the objects in the time domain so that there is no dramatic change in position or associated velocity between frames. The projection of a smooth three-dimensional trajectory is also smooth in the two-dimensional plane: and it implies that 'path coherence' is smooth and will not change abruptly. There are then three assumptions to make about the correspondence problem: the location of a given point: the scalar velocity: and the direction of motion will be relatively unchanged from one frame to the next.

But there have been many different approaches to hand tracking that need examination. Gonclaves et al. (1996) use a 3D model of an arm to track it against a

dark uncluttered background with the help of a recursive estimator. Gavrila and Davis (1996) also use a complex 3D model of the whole body and recover the 3D pose of the arm at each time step. In both cases the systems ignore the issues of clutter in natural environments. Deformable planar contours (snakes), coupled with Kalman filtering can be used for tracking non-rigid objects, like hands. Peterfreund (1999) improved the technique using optical-flow to detect and reject image measurements corresponding to image clutter and other objects. Hand contours were tracked in cluttered backgrounds by Isard and Blake's (1998) 'Condensation' technique, a statistical factored technique. But Heap and Hogg (1998) improved the technique to overcome discontinuities in contour shapes. The discontinuous shapes are described in terms of clusters in a high-dimensional shape space and the discontinuities are described in terms of transitions between these clusters using a learned Markov model.

Several blob-based methods have been used for tracking because they work when there are large changes between frames. Wren et al. (1997) tracked a person using a system that segmented the image into blobs using colour information. Prior information about skin colour and the topology of a person's body was used to interpret the blobs as a figure. Tracking human motion by grouping pixels with coherent motion, colour and temporal support into blobs, using the expectation-maximisation (EM) algorithm and Kalman filtering was undertaken by Bregler (1997).

Mammen et al, (2001), recognise that the articulated motion of the hand makes it very difficult to track while performing a gesture. Simultaneous tracking of both hands needs to deal with large inter-frame variations in shape, clutter and mutual occlusion. From this a model for tracking both hands in rectangular windows was implemented using skin-coloured blobs and the 'condensation' algorithm. A major problem in simultaneous tracking of both hands is hand-hand occlusion, so if there is significant overlap of the search windows, tracking is done jointly. A scheme was implemented that estimated when these erroneous observations occurred, based on the ones that were not erroneous, and the predicted values of the states. The technique was successful as it exploited the fact that, usually, not all measurements are occluded simultaneously and so it was successful in dealing with various kinds of clutter and occlusion.

Chen et al. (2003) use a very similar method for extracting features of skin-colour and motion as described in chapter 3. It is explained that the identified location of the hand will not necessarily be at the centre of the hand because the extracted information is often located on the boundary of the moving object. A refined centre point and hand outline is accomplished by a background subtracting and updating method. The motion analysis is used to characterise temporal features to be used as a feature vector for a HMM to recognise gesture.

The tracking technique described in this chapter is very similar to that described by Chen et al. (2003). Many tracking techniques require a prior model in order to deal with challenging local features. This method is capable of capturing non-rigid motion based on two powerful cues of colour and motion. Techniques are developed for segmentation of the combined colour and motion cue that allow a fair degree of latitude when setting segmentation parameters, especially in good lighting conditions. Furthermore, complete segmentation of the hand is found to be

unnecessary in many gesturing sequences and so the tracking approach is found to be a simple and robust method for this application.

The great attraction of the HMM is its ability to model temporal structure and variability. This technique was first used successfully in speech recognition. It uses a probabilistic pattern-matching approach which uses a time-sequence of speech patterns as the output of a stochastic or random process. Previously, dynamic time warping (DTW) had been used that took some account of time-scale variations (Owens, 1993).

But the technique used in this thesis, is a technique which has been used to transform data recorded at one sample rate to a different sample rate. One of the most widely used applications is to change music that has been recorded at 44.1 kHz, for CD (Compact Disc) and transform it to 48 kHz for DAT (Digital Audio Tape) (Ifeachor, 1993). This requires the number of samples to be increased by a ratio of 1.088. The number of samples in a sequence can only be altered by an integer amount. To produce an overall change in sample numbers that is not an integer value the number of samples must be first increased by interpolation. Then that new value is decreased by an integer amount by decimation. Calculations show that to produce the ratio of 1.088 the integer number for interpolation is 160 and for decimation is 147. This chapter describes a similar method used to adjust gesture samples to a normalised value of sixty-four.

Ellis (2002) states the importance of performance evaluation for video surveillance so as to assess the reliability of systems. These ideas can be transferred to gesture sequences so that algorithms or techniques can demonstrate robustness and correctness, to assess improvements resulting from incremental algorithm development. In addition, the use of widely available sequences provides an important opportunity for benchmarking. Alternatively, another method is to assess algorithmic performance using synthetic image sequences.

4.3. Data generated from gesture sequences

Poor segmentation of the hand needs to be considered. With reference to the previous chapter, a technique for locating the hand position was described using skin-colour masks and motion masks to produce gesture objects, in a similar way to Chen et al. (2003). The resulting objects are sorted by area into rank order so that the largest area has a higher likelihood of being the most significant skin-coloured object in motion. There is the case where a large skin-coloured object, like the face, may make a small movement and be of comparable size to a relatively smaller area, like the hand, moving just a small distance. However, the spatial location of these objects is usually so different that an object selection algorithm is able to discriminate between the objects as to whether they are hand, face or clutter, unless the hand passes in front of the face.

The direction and distance moved by the hand from sample to sample can vary widely, so any prediction of the next position can be very unreliable. Even a small amount of averaging to remove noise from the data can affect prediction calculations, especially at a turning point of the trajectory. At normal sample rates of 25 frames per second there is often insufficient data to perform filtering before a change of

direction takes place. But the impulse response of a Kalman filter takes a significant number of samples to form its current estimate (Lynn, 1985) and hence this approach is not used.

In an ideal situation there would be just three rank-ordered gesture objects to locate and record associated with the hands and the face. However, the fragmentation of some skin-coloured and motion regions produce more than three objects and can be further exacerbated by the generation of additional noise type of objects. It has been found useful to record the first six rank ordered objects even though the first two or three are usually all that is required for the tracking process. Figure 4.1 shows a more ideal type of result in which there is only movement from the dominant gesturing hand. The left hand image shows the centre of gravity of the only gesture object. The mask on the right image shows the silhouette of the hand. This object has been produced by finding the coincidence of the gesture object with the hand region, defined by the Hue mask. Assigning gesture objects to Hue regions is to produce assigned or SCMI objects.

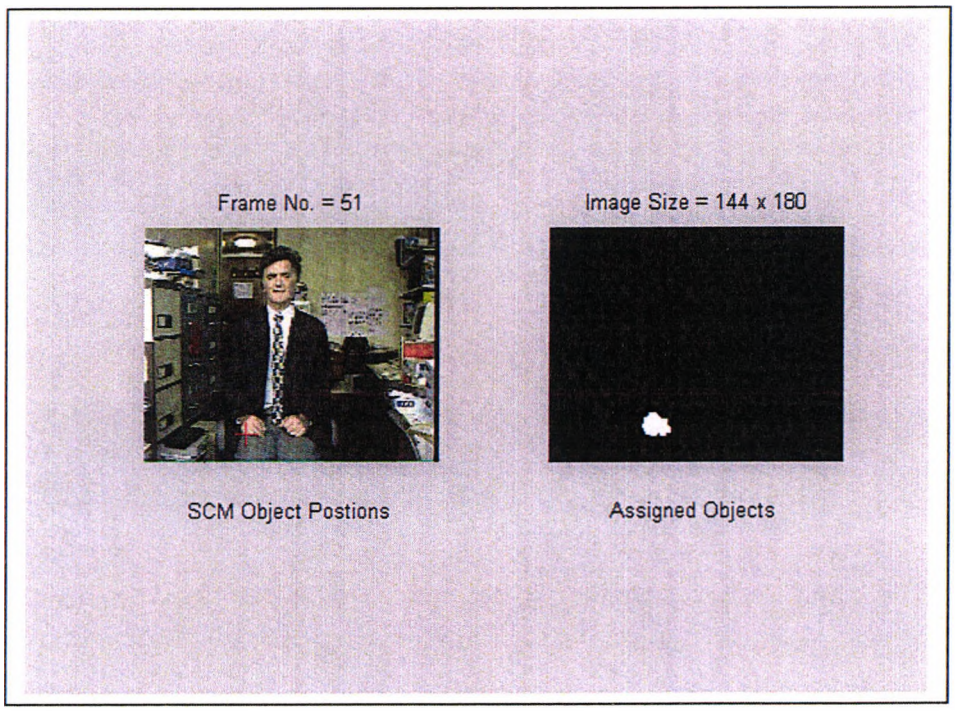


Figure 4.1 Gesture object shown by red cross (left) and the assigned or SCMI object (right) showing the hand outline.

Figure 4.2 shows the position of the first three most significant gesture objects, in 2D and 2DT views over a sequence of fifty-one frames. The first or most significant object is labelled red and the second and third objects are labelled green and blue respectively. The gesture sequence consists of the right hand moving to the left shoulder, tapping it and then returning to the original position (a gesture in the Makaton language for indicating that a visit to the toilet is required). The 2D and 2DT images of Figure 4.2 show the distribution of the three most significant objects. It can be seen that the three objects are generally associated in the same space as each other, as a result of being generated by different parts of the same hand. There are indications that some objects have originated from other, non-human parts of the image in the image sequence. The selection of which objects are relevant for tracking

are determined by an algorithm termed the Object Selection Algorithm as it is implemented after the rank ordering of objects. This is discussed later in this chapter.

When the SCMI objects are recorded, in this environmental situation, the number of objects is much reduced. This is shown in the 2D and 2DT images for the sequences shown in Figure 4.3. The most significant red object is seen to describe most of the trajectory. The instances of the green and blue objects making continuous lines, in the 3D image, is because the plotting algorithm repeats the last known coordinate when no gesture object is generated at a particular frame. This can be due to only the first significant object being generated; poor segmentation or no additional movements.

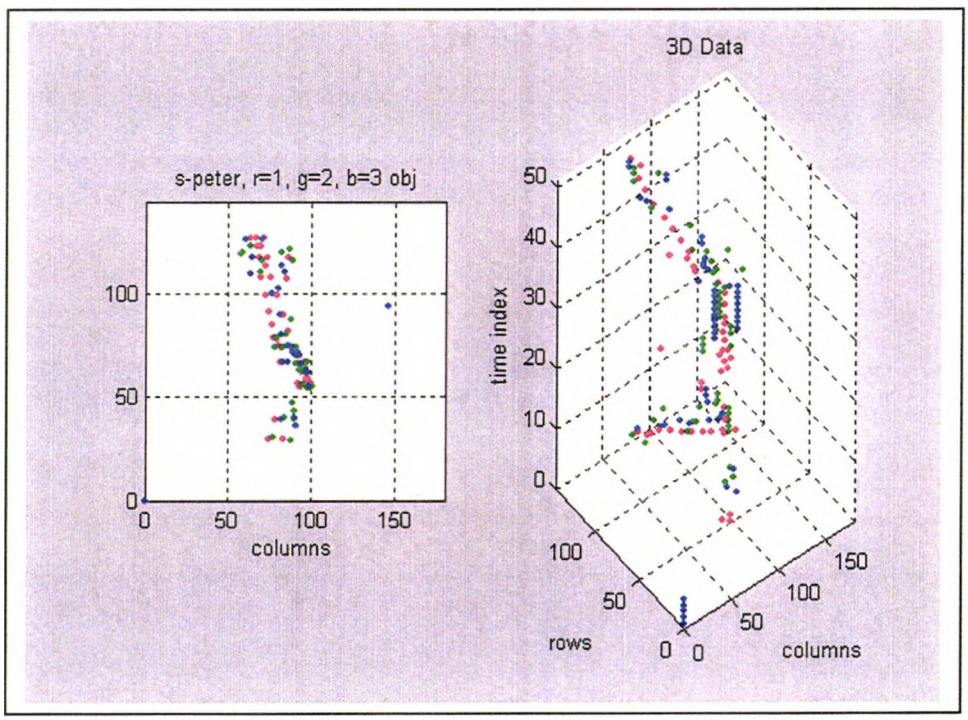


Figure 4.2 A sequence of 51 frames (time index) showing 2D (left) and 2DT (right) images of the positions of the three most significant rank ordered SCM objects (red, green and blue, respectively).

Figure 4.3 clearly shows that the generation of SCMI objects produces a reduction in the number of gesture objects. The most significant object (in red), as shown in the 2DT image, tracks most of the trajectory of the hand. The straight lines of the second most significant object (green) and third most significant object (blue) indicate that they have not been generated for most of the trajectory and that the most significant object is the only object generated for most of the sequence, the exception tending to be near the start and end of the sequence. The multiple objects in the previous figure are due to fragmentation of the gesture objects in the hand region which have now been mainly reduced to one object. The potential advantage of using the SCMI object is that for the majority of the sequence only the one object is generated. The object relates to the dominant hand position and so tracking is essentially just tracking the most significant object.

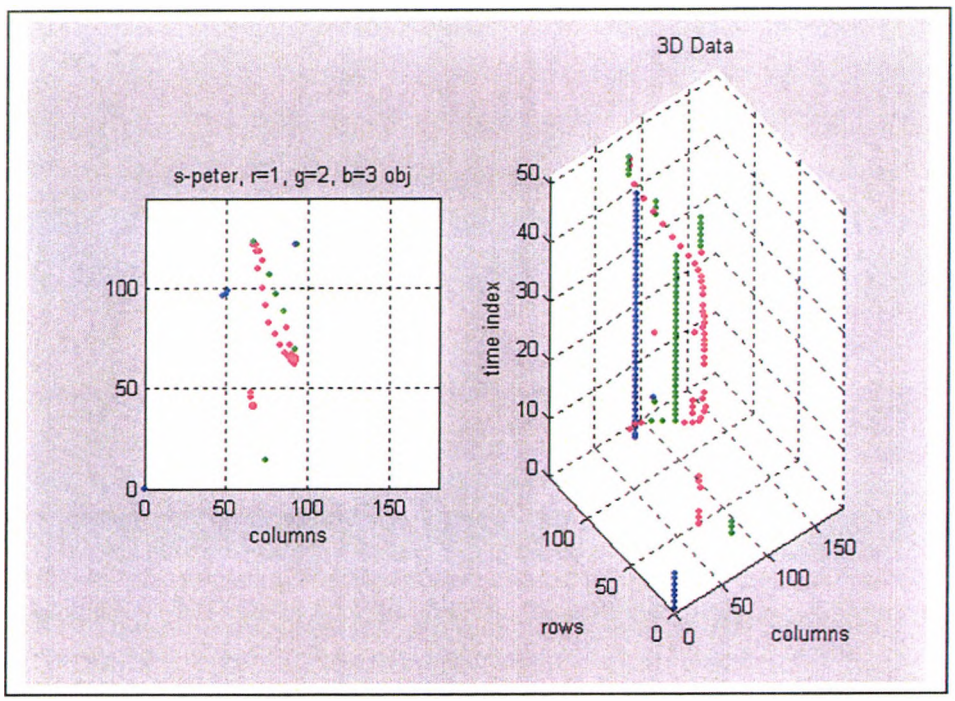


Figure 4.3 A sequence of 51 frames (time index) showing 2D (left) and 2DT (right) images of the first three most significant rank ordered SCMI objects (red, green and blue, respectively).

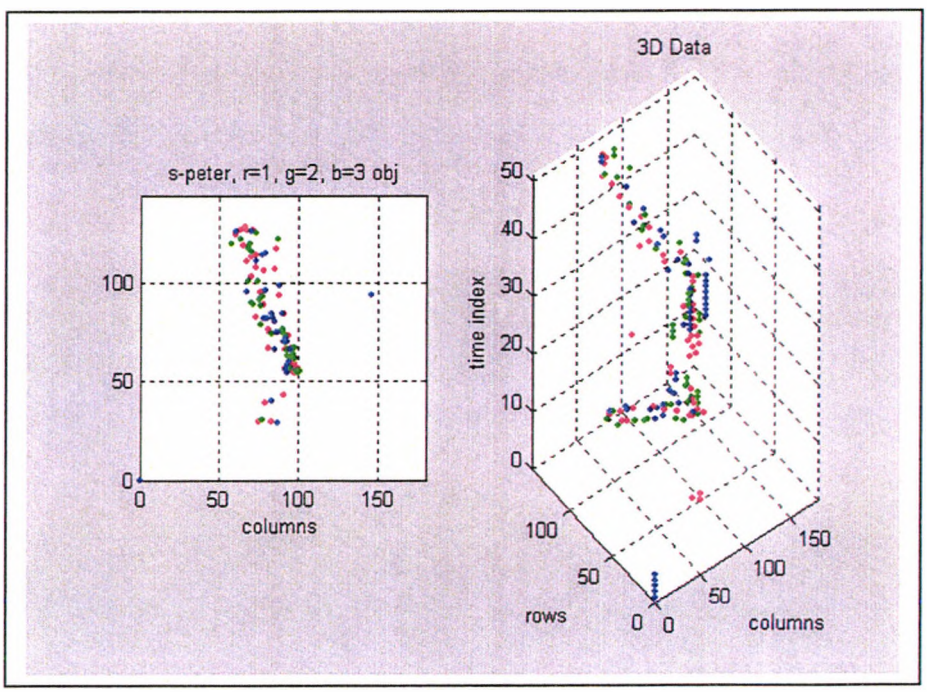


Figure 4.4 2D (left) and 2DT (right) images of the first three most significant rank ordered SCME objects (red, green and blue, respectively).

A more profound result was recorded when incorporating the coincidence of the finger region with skin colour and motion regions by the logical AND of the motion, skin-colour and edge masks to produced SCME objects. These objects give rise to a distribution, as shown in Figure 4.4, of the first three significant ranked objects not

much dissimilar those shown for SCM objects in Figure 4.2. However, assigning the objects to the skin-colour mask to produce SCMEI objects shows a dramatic reduction in the generation of objects and the most significant object, as shown in Figure 4.5, tracks mainly the hand. This process shows that the first object will suffice in characterising the trajectory and that the second and third objects are virtually redundant.

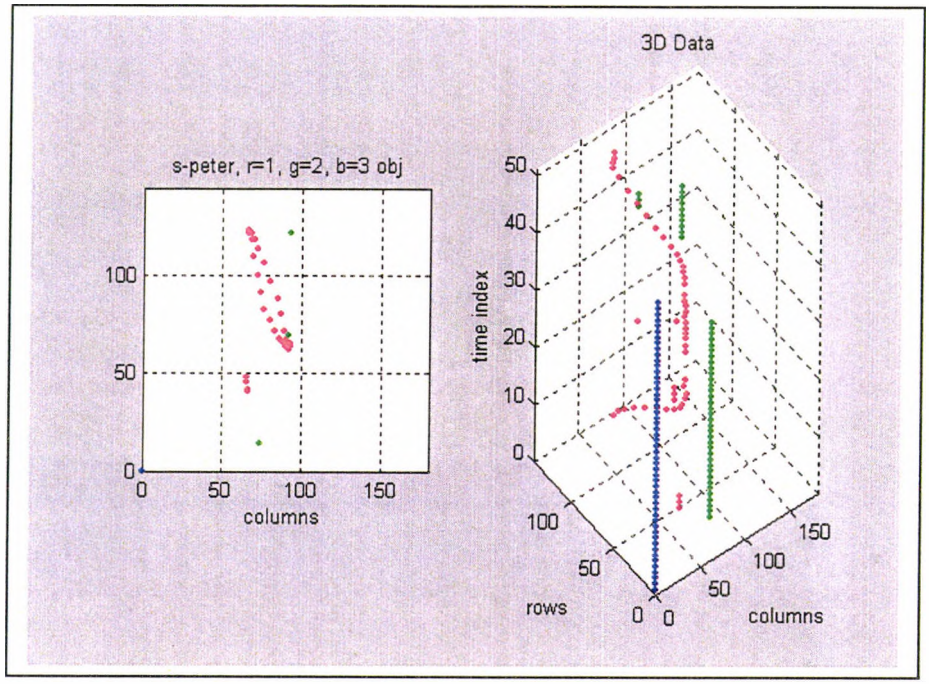


Figure 4.5 2D (left) and 2DT (right) images of the first three most significant rank ordered SCMEI objects (red, green and blue, respectively.

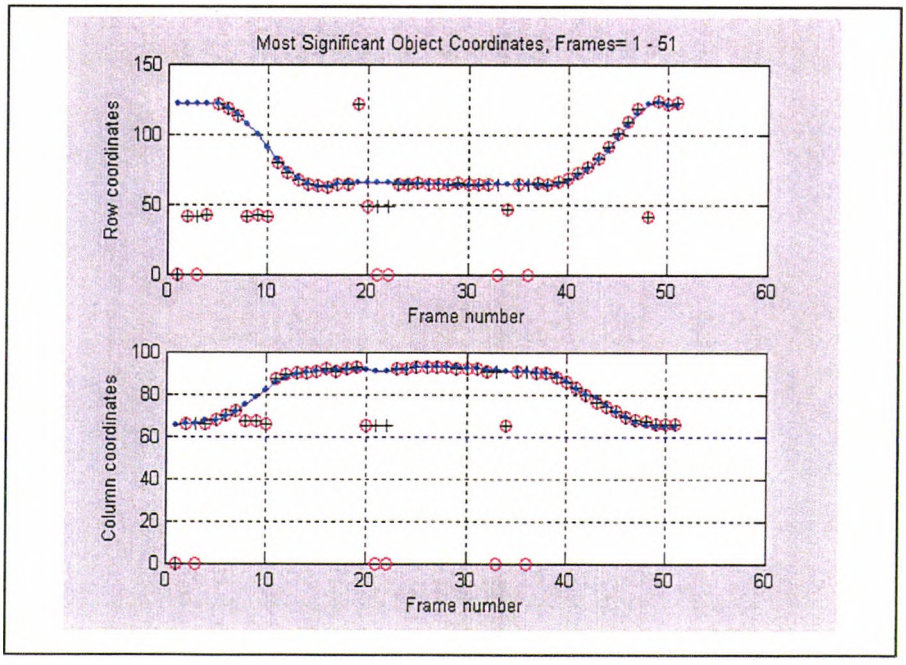


Figure 4.6 Comparison of the most significant data (red 'o') with manually obtained data (blue '.') and output position (black '+').

The SCMEI data is shown in a different format in Figure 4.6, in which the row and column data are shown separately, with red circles indicating the coordinates of the most significant object. The blue line in Figure 4.6 indicates the approximate centre of gravity of the right hand that was recorded manually/visually, for the whole sequence. It can be seen that the coordinates of the SCMEI data closely coincides with the visual data for much of the sequence. There are frames when the hand position is not shown by the most significant object. For example, at frames 8, 9 and 10 the most significant object is due to head movement and at frame 19 it is due to the left hand movement. At frame 21 and frame 22 there is an example of no movement being detected so the coordinate data is zero. The black crosses indicate the output trajectory coordinates. When a static condition is encountered the output takes the previous coordinate position. However, this simple correction algorithm is limited in its use, as inappropriate object coordinates (head) can be recorded as seen in the column data of frames 8,9,10, 20 and 34.

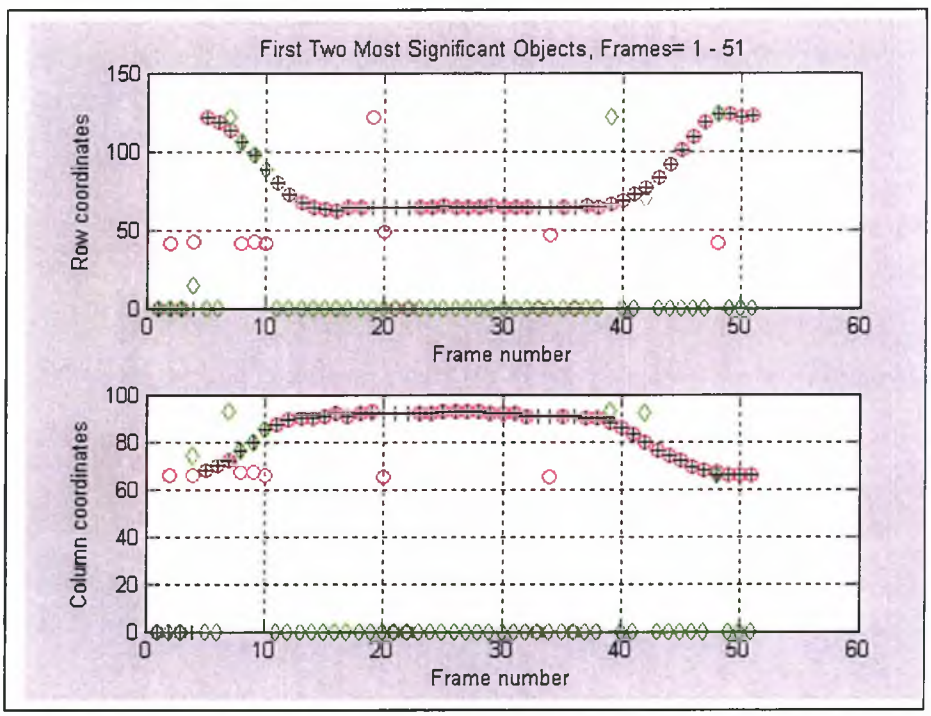


Figure 4.7 The inclusion of the second most significant object (green diamond) improves the tracking performance output (black cross).

The inclusion of the second most significant object in the tracking algorithm shows an improvement in the resulting trajectory. This is shown by the black crosses of Figure 4.7 for frames 8, 9 and 10. More significantly, incorporating criteria for detecting a sudden change from the previous hand position, improves the tracking performance. For instance the change of coordinate value at frame 19 and frame 20 was detected. The coordinates at frame 20 were substituted by the previous value (frame 18), a more likely value.

4.4. The complete OSA (Object Selection Algorithm)

The rank ordering of objects whether they are SCM, SCMI, SCME or SCMEI objects increases the likelihood that the most significant object will normally represent the position of the dominant gesturing hand. However, the previous section has shown that this process is not perfect. There are occasions when other factors have to be considered. Whilst the gesture is in its dynamic phase the criteria that the highest ranked object, by area, locates the position of the moving hand successfully. However, when the hand becomes static or 'partially-static' because of no or little movement, other movements, typically the non-dominant hand or the face, can make a more pronounced movement. This movement can mean that the dominant hand may become the second or third highest-ranked object. The problem then, is to decide which object represents the dominant hand.

Another problem, probably due to the recording process, occurs in the dynamic phase of the gesture in which adjacent frames are the same: the same images do not produce any motion detection and data is blank for that pair of images. The algorithm has been designed to fill in missing data using the same data as the previous data. Prediction or interpolation routines have not been found necessary as the movement in these situations would be minimal. There is very little error in using the previous coordinate values. Furthermore, as will be discussed in the next chapter, when considering the frequency content of the gesture trajectories, small variations in location of the hand are not important and can be regarded as a high frequency signal. This does not affect the characteristics of the gesture.

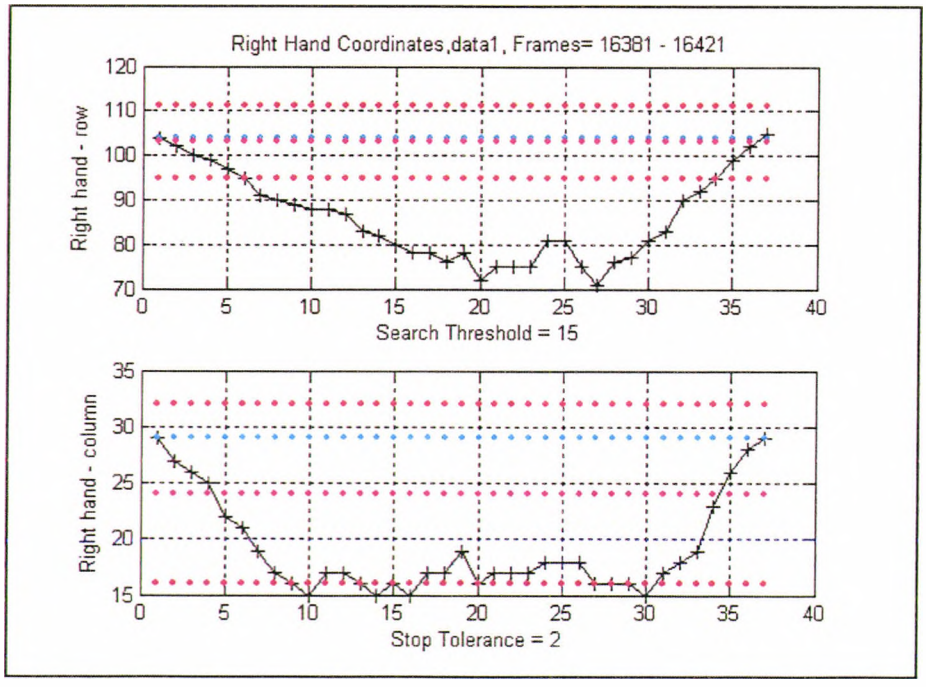


Figure 4.8 Initial coordinates shown by magenta dotted line and the red dotted line showing the upper and lower tolerance to this estimate. Stopping criteria set by the updated initial position as shown by the cyan coloured line.

The Object Selection Algorithm (OSA) was initially developed for single-handed gestures, but some opportunities were found to incorporate tracking of both hands. Data is first fed to the algorithm giving the likely initial coordinates of the right and/or left hand. A window or range of coordinate values is allowed for because of the difference in coordinate values from the static hand position to the first detected movement. The hand coordinate position is then updated at detection of the first movement. This allows stopping criteria to be set and activated to determine when the dominant hand returns to the initial position at the end of the gesture. At this stage a start flag is set to determine where the gesture starts, which is defined as the first movement away from the initial resting position. Figure 4.8 shows a typical trajectory with the magenta dotted line showing the initial position estimate, and the red dotted line the upper and lower tolerance to this estimate. The dotted cyan line shows the value of the initial updated position of the right hand, which is used as stopping criteria for recording the gesture. The initialisation routine to set up the initial trajectory coordinates is shown in Figure 4.9.

In the first instance, the algorithm will determine if the first or second object is closest to the initial condition. If it fails, then the algorithm then passes to the next frame and tests the new set of data until the condition is met. It has not been found necessary to extend the search to the third most significant data as no overwhelming case has been found for its inclusion, as yet. However, it could be included if needed, but at the cost of greater complexity in the algorithm.

After initialisation, objects are tracked according to the algorithm detailed in Figure 4.10. If the dominant (normally right) hand coordinates are known, from the last frame, then tests are required to determine the suitability of the current data. The first test is to determine if the first (most significant) object is close to the previous object. A tolerance value is set that sets the maximum distance that the two instances can be apart. If the test is affirmative then the object coordinates are assigned as the next position. If it fails, then this can be due to the non-dominant (normally left) hand, face or some other object being detected. If two-hand tracking is being undertaken and it is in the vicinity of the left hand then it can be assigned to the next position of the left hand, otherwise, for single hand tracking, it is discarded. Two hand tracking is detailed in Figure 4.11.

If the most significant object has been discarded in the OSA then the second most significant object is tested for its suitability of representing the hand position. The second object is tested for its closeness to the first object, by a tolerance related to the hand's dimensions, which results when there is fragmentation of the gesture objects in the hand region. In this case the second object position is ignored and the first object's position is used. The second object is then tested to see if it represents the position of the dominant hand, by using the same distance tolerance as was used for the first object. If the test is affirmative the second object's location is assigned to the trajectory location, otherwise the second object is ignored. If the first or second objects are rejected as not suitable for the current coordinates then the coordinates take the previous coordinate values. This has not found to distort the gesture trajectory, as the consideration of objects other than the most significant object occurs at times of little movement. Previous values fill the gap in results with minimal error in the trajectory shape.

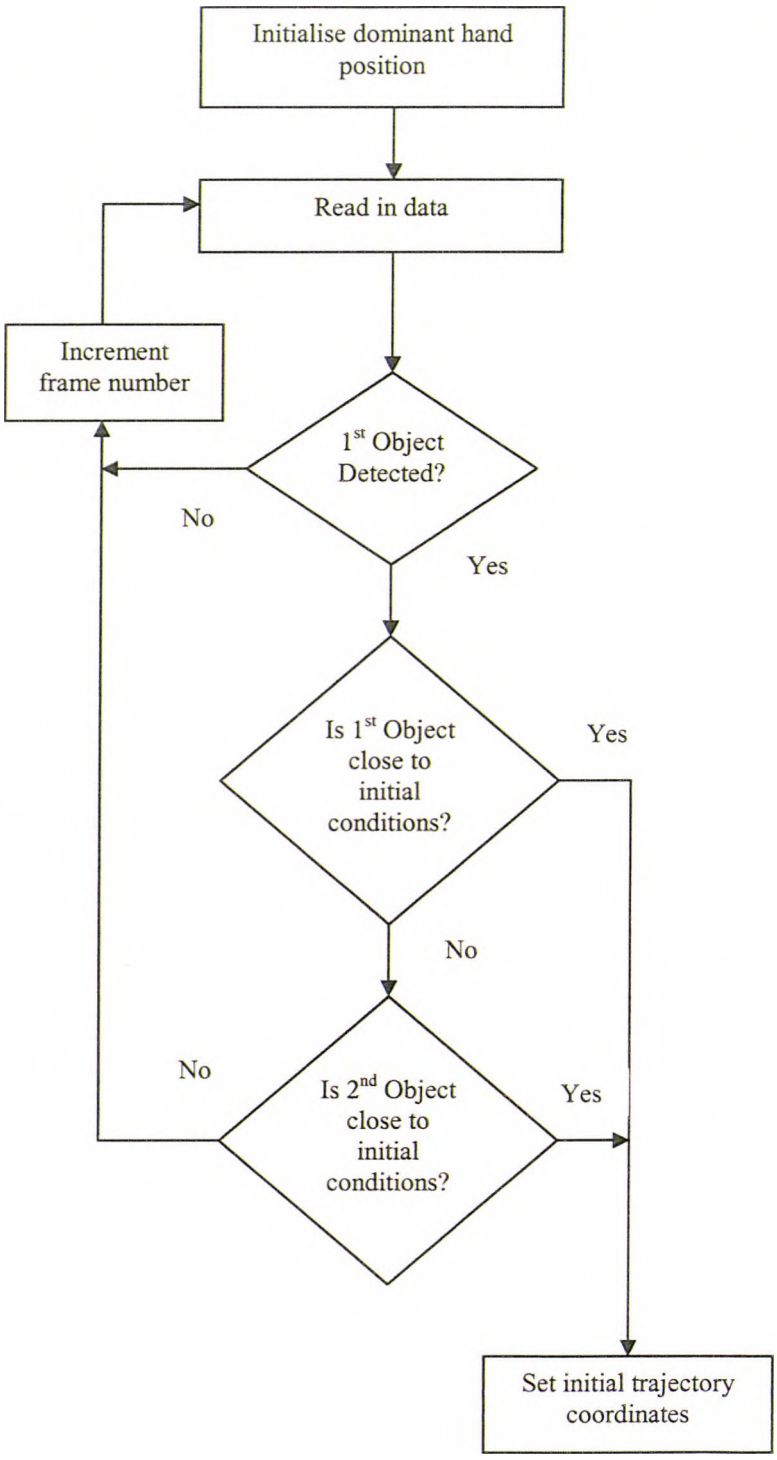


Figure 4.9 Initialisation of the OSA to set initial trajectory coordinates

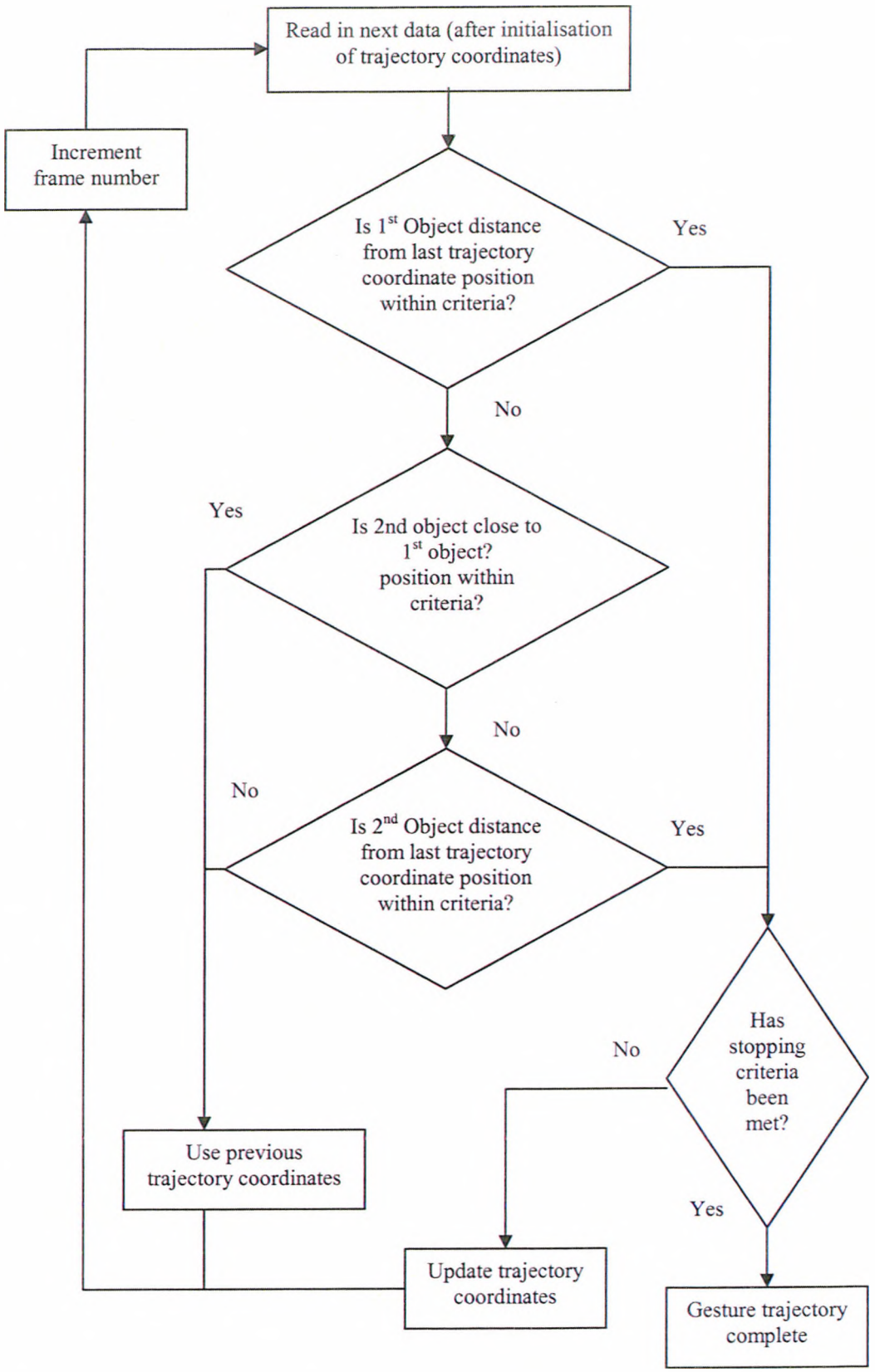


Figure 4.10 OSA using just two gesture objects to follow the dominant hand

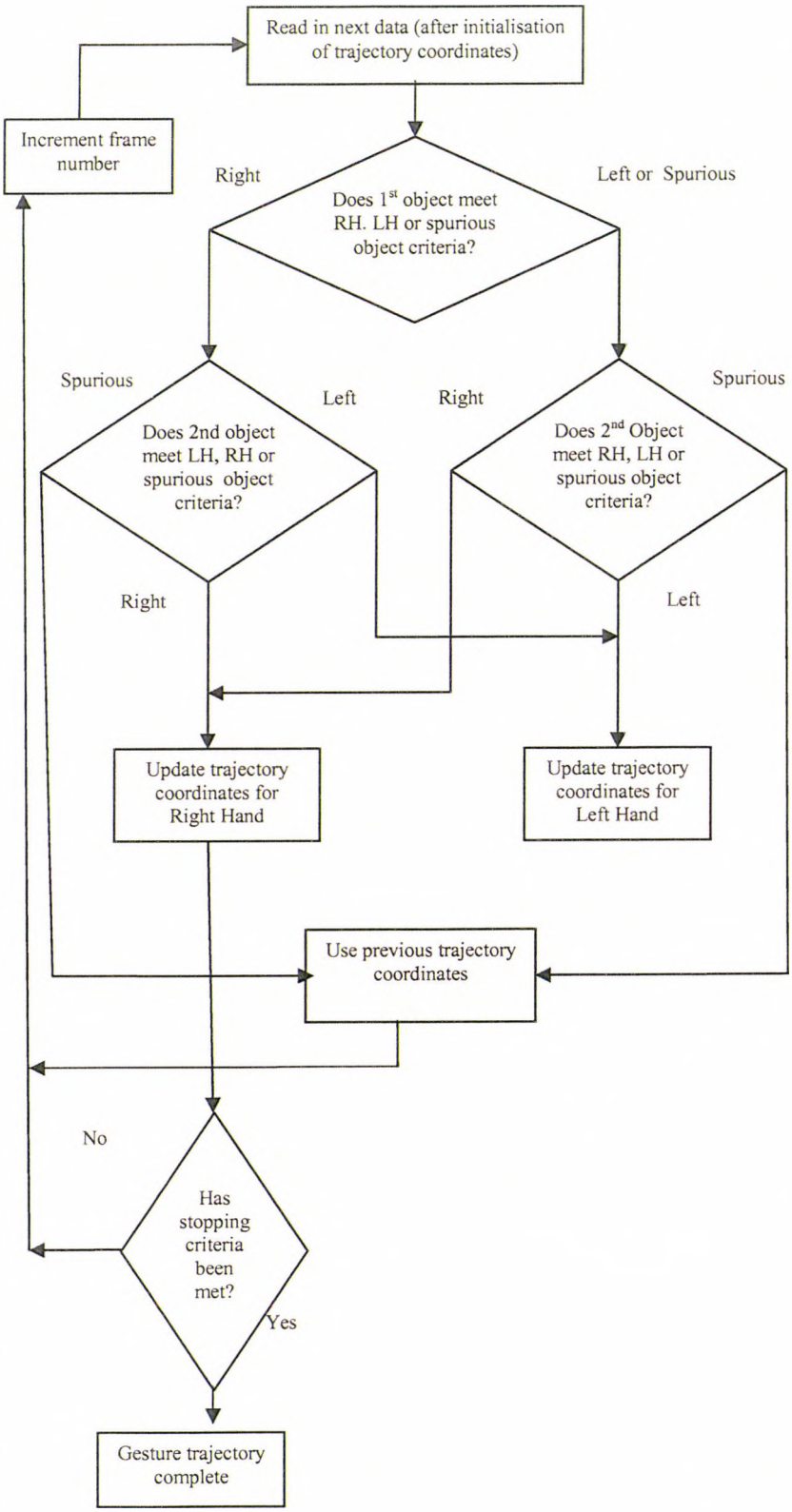


Figure 4.11 OSA using just two gesture objects to follow both hands

The gesture trajectory end is normally determined when the hand position returns to the initial updated starting location. As this position is not always exactly the same, some latitude is allowed and a tolerance is included. A full investigation into the stopping tolerance and its effect on gesture results is analysed in Chapter 5. Other stopping criteria are included if zero data is recorded and excessive constant data is recorded, indicating a static mode. The gesture data is then ready for the time normalisation process.

4.5. Choice of Sequence Parameters

The performance of the OSA using the SCMI objects showed several deficiencies when sequences were recorded in poor lighting conditions, as the hand shape became less well segmented. This also affects the results when edge detection was incorporated into the mask and resulted in loss of data in the sequence. However, the SCM objects produce data that related to the hand position and produced reliable results with the OSA.

Figure 4.12 An image from a sequence recorded in low illumination and poor white balance

A range of tests was conducted in five different environmental conditions as described in Appendix II. In some of these tests there was just a single gesturer performing a single gesture. In other sequences there were a number of different gesturers who performed a range of gestures, one after another. In addition more than one person was in the image with one of the people performing the gesture several times. An example of one of the images from a poorly lit 250-frame sequence is shown in Figure 4.12. This sequence is lit by fluorescent light and natural light illuminates the left of the subject. The image could be considered to be dark.

In the previous chapter the choice of using just Hue or Hue and Saturation masks for the segmentation of skin-coloured objects was discussed. Experimentation on sequences, taken in the environmental conditions described in Appendix II gave the

impression that as lighting became less uniform, data was not so reliable or missing. Furthermore, SCMI or SCME objects were more likely to be unrepresentative of the hand region because of poorer skin-colour segmentation and the threshold levels associated with edge detection would need constant adjustment and the process was likely to fail. Additional parameters were investigated to find out if noise removal or filling in holes in final binary mask from which the SCM and SCMI type objects are produced might improve reliable object generation. The morphological operations of 'opening' or 'hole filling' were applied to the sequences. The opening is used to remove the borders of frayed regions and for eliminating tiny regions: -

$$f \circ s = (f \ominus s) \oplus s$$

where the binary image 'f' is eroded first by a structuring element 's' and the resultant image is then dilated by the structuring element. Whereas, hole filling changes the value background pixels surrounded by foreground pixels from 0 to 1.

The details of a typical poorly lit sequence and experiment are shown in Appendix III. Data was recorded for the eight conditions using Hue or Hue and Saturation for determining skin-colour range and with and without 'opening' and hole-filling, and four types of data i.e. SCM data ; SCMI data; SCME data; and SCMEI data.

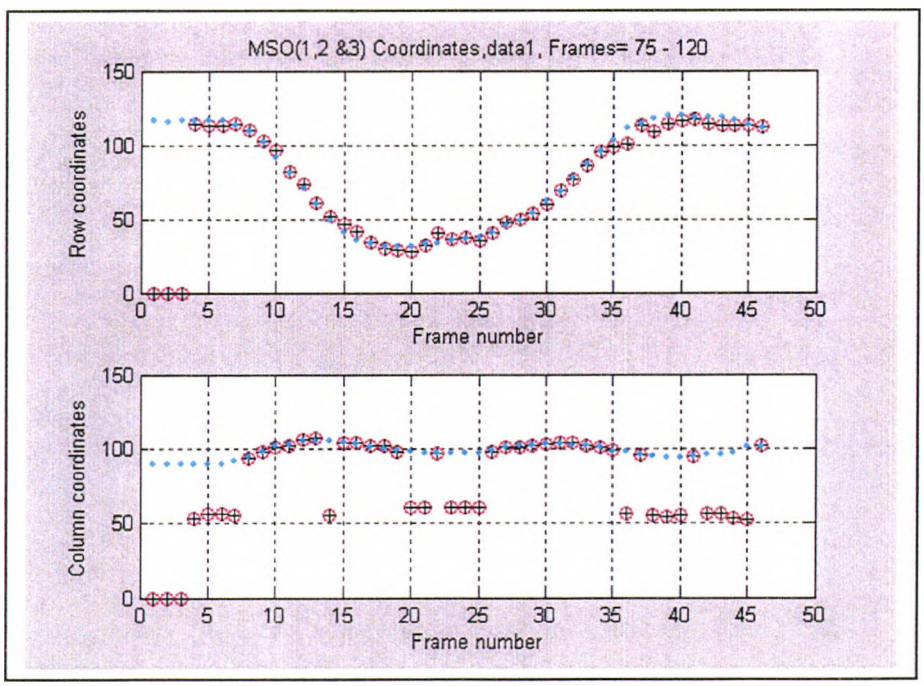


Figure 4.13 The most significant SCM object for the two hand sequence with the left hand visually obtained data (cyan) and the most significant object coordinates (red).

The hand coordinate data in Figure 4.13 shows the distribution of the most significant SCM object for both hands waving. The conditions that produced this data were for no hole filling and no opening and for the saturation range set at its default range of 0.05 to 0.95. The column data shows the distribution of the data between the right and left hands. The positions of the row data of the right and left hand coordinates are difficult to separate, as they are virtually identical. When the experiment was run again but with Saturation changed to a more restricted range, for

better segmentation of the skin-colour regions from the background, very little difference in results were obtained. The row data was virtually identical, whereas there was some change to the distribution of the most significant object to the right and left hands. In the former case, 25 objects were assigned to the right hand and 18 to the left hand, and in the latter case it was 29 and 14 respectively. The cyan coloured dots show the visually recorded data of the trajectory of the left hand and show close agreement with the automatically produced data.

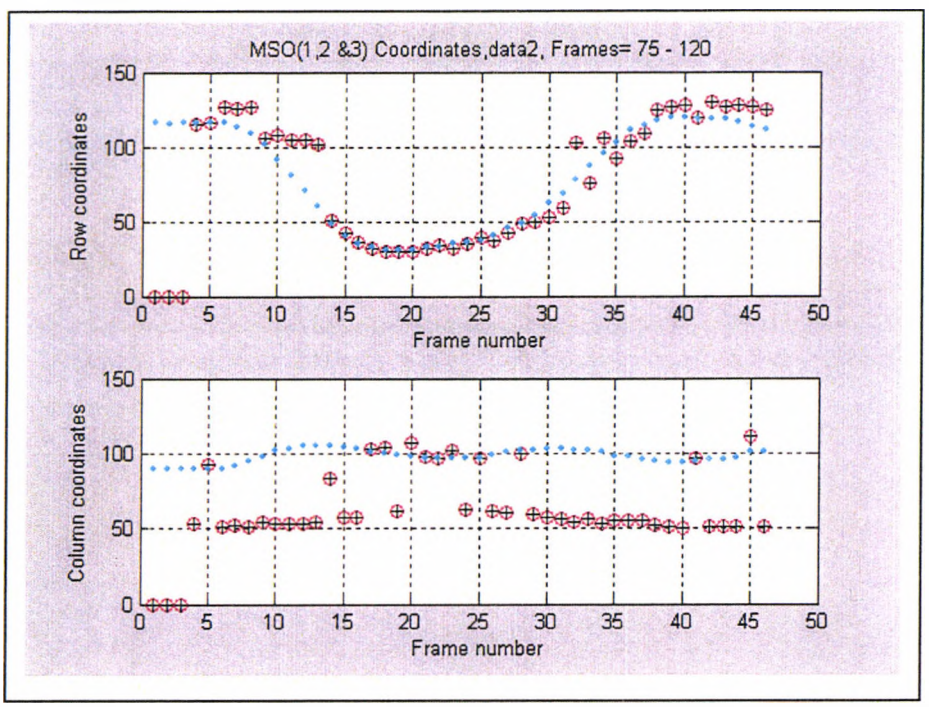


Figure 4.14 The most significant SCMI object for the two hand sequence for the optimum Hue and Saturation range with visually/manually obtained data (left hand, cyan) and the most significant object coordinates (right hand, red).

The SCMI data was very poor when the saturation was at its default range. This was to be expected as segmentation included many background features. Improvements to the data quality were made with a more appropriate saturation range, as shown in figure 4.14, but there are significant regions where object coordinates do not represent the expected trajectory, as shown for example by the cyan coloured locations of the visually recorded data. The edge detecting technique also failed to work very well in these conditions with little coordinate data being generated.

The effect of using hole-filling or opening was minimal except close to static conditions. In the case when the optimum Hue and Saturation ranges were selected and hole-filling and opening was applied, no data was recorded at frames 22 and 23. This is due to the small overlap between the skin-colour mask and the motion mask being removed by opening. In other words the sensitivity to detecting small motion has been lost.

The inclusion of the second most significant SCM coordinates into the data stream show (Figure 4.15) that two hands trajectories can be obtained by a combination of the first two SCM objects. In this figure the visually obtained trajectory of the right hand is shown by a series of magenta dots and the left hand by the cyan dots. Green diamond shapes show the second most significant data coordinates. It can be seen

that for a given trajectory tracking follows a sequence of first and second objects. The tracking of both hands and the comparison of ground truth data for both hands is shown in Appendix III. Tracking of both hands failed when the distance between the pair of hands fell below the search threshold, typically when the hands go into a crossing trajectory.

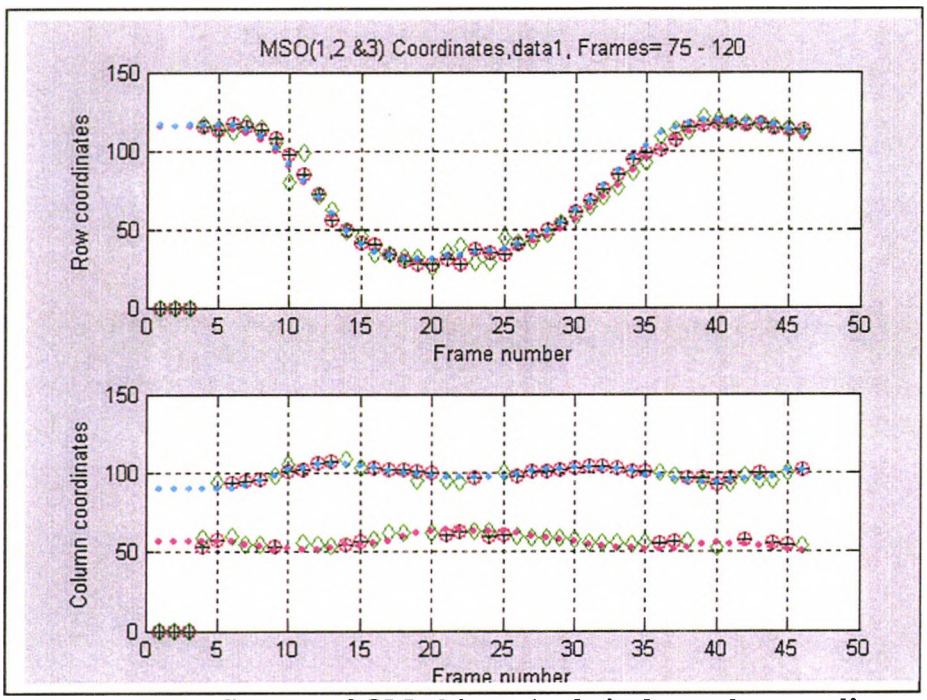


Figure 4.15 The first two SCM objects (red circles and green diamonds, respectively) coordinate identify the trajectory of the right and left hand, with cyan and magenta dotted lines showing the visually obtained left and right hand ground-truth data respectively.

4.6. Three people in an image

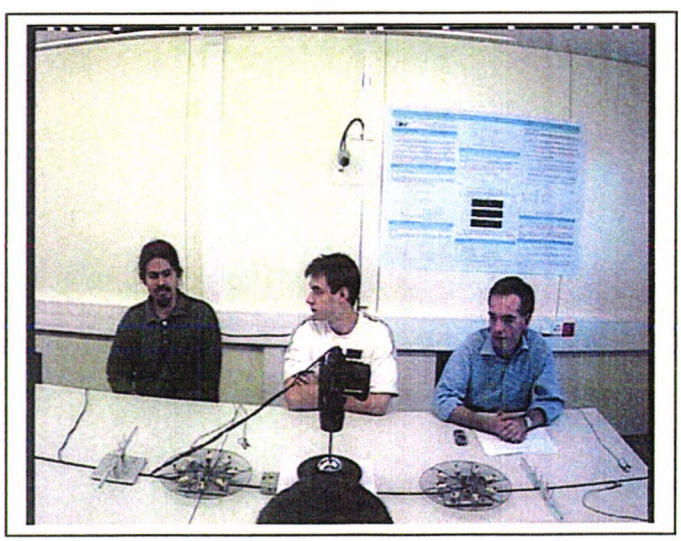


Figure 4.16 Three people in a PETS image

The original hypothesis for the generation of SCM objects was for the case of one person in the image and three moving skin-coloured objects in the image. Subsequent experimentation showed that in poorly lit situations, or at near static conditions there was fragmentation of the SCM objects. Moreover, the uncovering of a skin-coloured background object could appear momentarily as a SCM object. However, tracking could still be undertaken successfully and for most situations the first two SCM objects were sufficient.

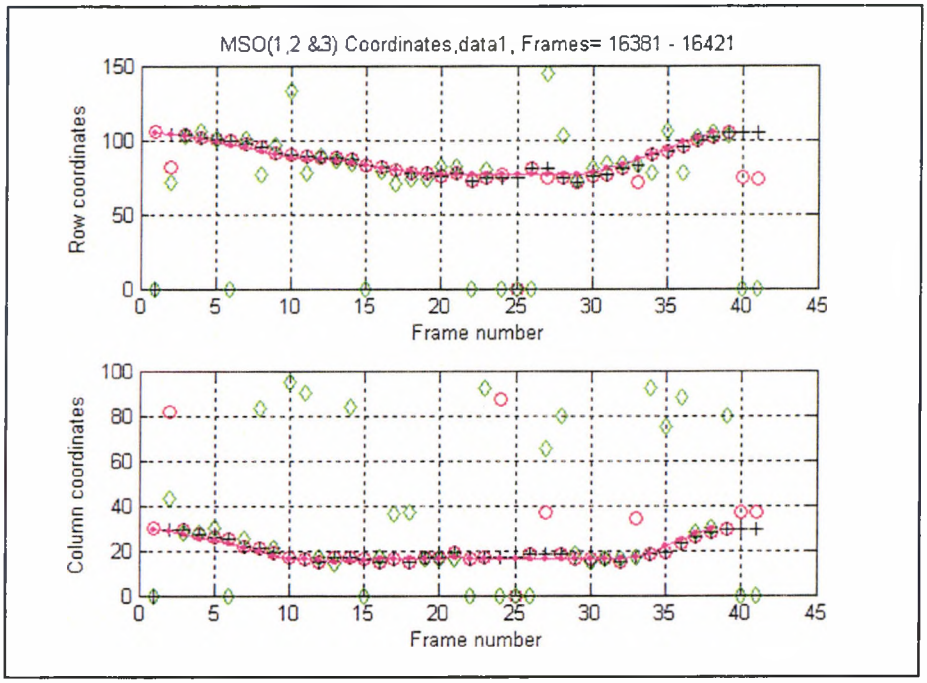


Figure 4.17 Tracking of PETS data with a '+' from first two SCM object coordinates (red circles and green diamonds, respectively) and comparison with visual data (magenta)

The use of publicly available sequences of people gesturing were considered for investigation. The PETS (Performance and Evaluation of Tracking Systems) database uses three people in the scene, as shown in Figure 4.16. The main gesturer was the person on the left of the image. Although the two other subjects were not deliberately gesturing and the image could have been cropped to exclude them from the analysis. However, it was decided to apply the techniques to the whole image even though movement of the non-gesturers could be observed.

In a part of the image sequence the person on the left gestures by raising his right hand, the other two people are also seen to move. A full analysis for a 48 frame sequence is tabulated in Appendix III, where the first three SCM objects are shown to mainly relate to any of the skin-coloured regions of the three people. The two people to the right of the main gesturer (person on the left in the image) are not stationary and their movements are generated as SCM objects. For instance, at frame 16967, (Figure A4.3) about the eight from the start of the sequence, motion is detected on the face of the left hand person; the right hand person and the centre person, in that order as tabulated in Table A3.3. The result is that the data is more sparse than would have been expected with just one gesturer in the image, but the

algorithm was able to successfully track the hand position of the person to the left of the image, regardless. The tracking of the right hand of the left-hand person is shown in figure 4.17.

4.7. OSA Performance

The difference between the ground truth data i.e. manual location of the hand and the position of the hand generated by the Object Selection Algorithm from SCM objects give interesting insight to the algorithm's performance. The following example is taken from the two handed tracking experiment as shown in Figure 4.18. Right-hand tracking is undertaken with the ground-truth data shown as the magenta line. The OSA's output is indicated by the black '+' symbol, which in this example seems to follow the second most significant SCM object more than the most significant SCM object as that is favoured by the left-hand more frequently.

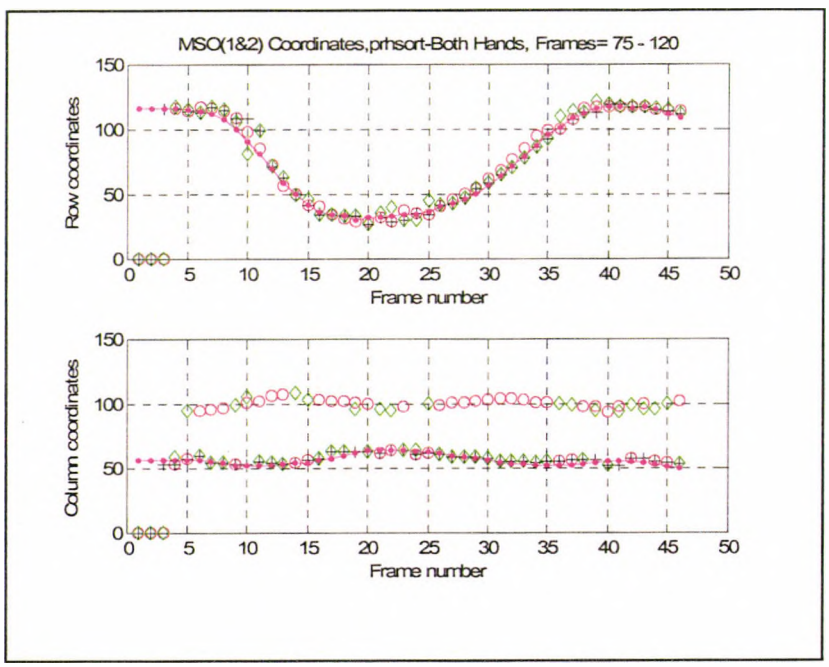


Figure 4.18 An example of tracking data to access the OSA's performance.

Initially, the bar chart in Figure 4.19, shows a large difference (frames 1 and 2), caused by the manual data locating the hand coordinates but the OSA data is at zero as no motion has occurred. As soon as motion of the hand is detected, SCM objects are generated and the OSA determines which of the first two most significant objects to track. It is interesting to note that the difference value is quite low and is usually much less than the length of the hand (18 pixels from wrist to finger tops) which is shown as the distance between the two red lines on Figure 4.19.

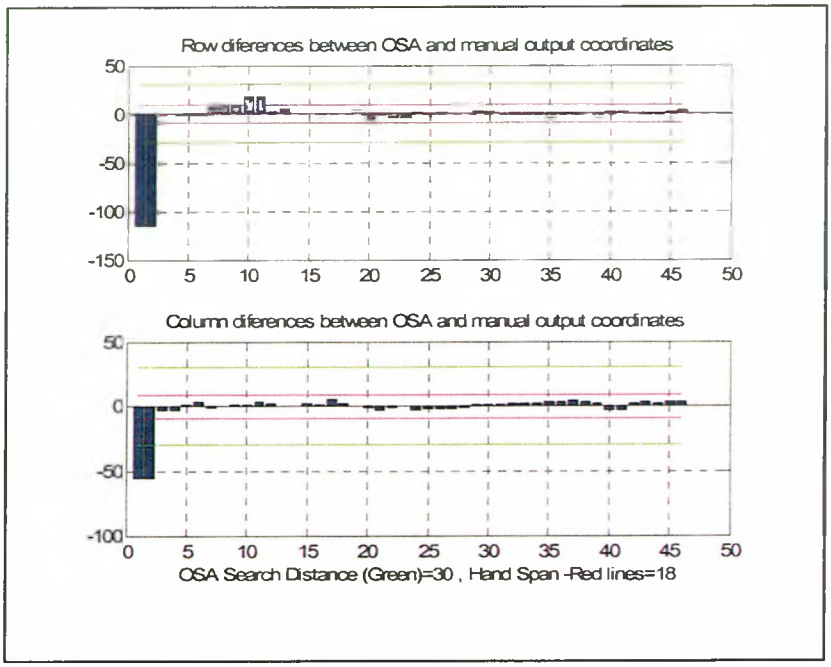


Figure 4.19 Bar charts showing difference between manual and OSA for row (upper) and column (lower) coordinates. The hand span is shown between the red lines at 18 pixels and the search window set at 30 pixels (green lines).

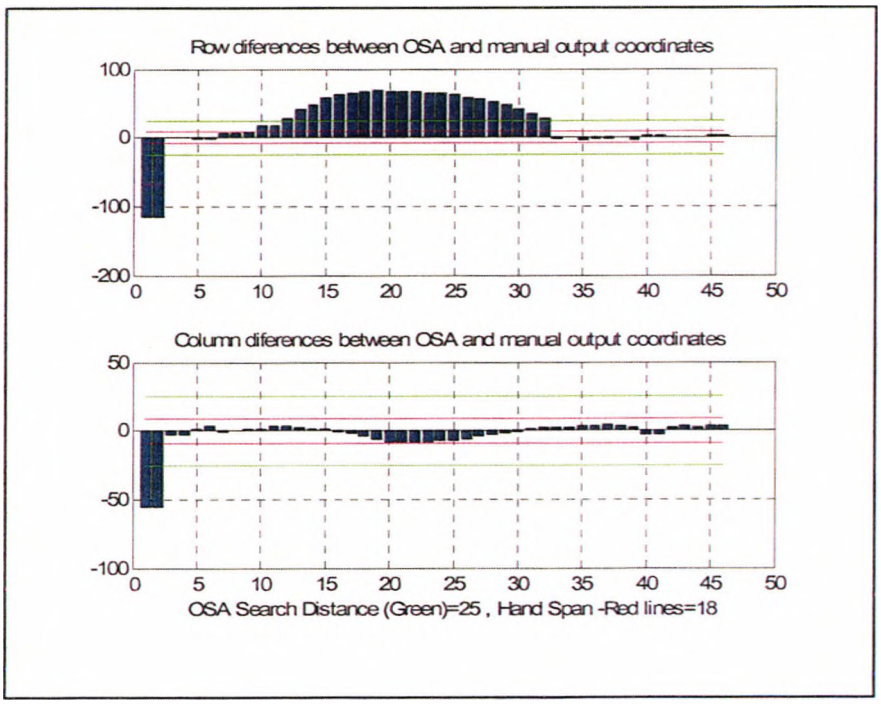


Figure 4.20 Search distance insufficient for good tracking

In this example there are two occasions when the difference is greater than the span of the hand. This occurs at frame 10 and 11. At frame 10 the first two SCM objects are both related to the left hand and the third SCM object is related to the right hand, but not used in the OSA. When this occurs the OSA substitutes with the last coordinate values for the right hand. This normally causes little error in the tracking procedure. However, in this case the hand is entering the stroke phase and the distance covered between each frame is a maximum. In this situation it is important

to adjust the search distance, shown by the green lines on the figures, so that the larger movements can be accommodated. Figure 4.20 shows the situation when the search distance has been reduced from 30 to 25 and the tracking procedure fails from frames 9 to 33, and then completes the tracking to the end of the gesture at frame 46. During the tracking failure the bar graph of figure 4.18 shows significant differences that are greater than the hand length, particularly in the row coordinate figure. It is interesting to note that in this particular example the column differences are within limits of hand span, but the row differences greatly exceed the threshold, indicating much greater vertical than lateral movement

4.8. Time Normalisation

The appearance-based tracking data generates spatial co-ordinate data and time-index data relating to the frame number for every frame in the sequence. The spatial (2D) view and the spatial-time (2DT) view are shown in Figure 4.21. Instead of equally spaced boundary samples of an object, this application uses equally spaced time steps of the co-ordinates of the trajectory of the hand, which are the centroids of the SCM or SCMI object selected by the OSA.

For object recognition, using the Fourier Descriptor technique, the steps around the perimeter form a closed space, returning to the start point. In this application, the steps around the perimeter are replaced by time steps of coordinate data that represent the hand position trajectory and return to the same or near the initial starting coordinates. In sampling the perimeter of an object as, for example, when using Fourier Descriptors the starting and stopping coordinates are never exactly the same, as this would invalidate the condition of periodicity. In addition, a gesture trajectory will exhibit a finite difference between the coordinate values of the static starting position and the first detected movement coordinates as well as the starting and final stopping coordinates.

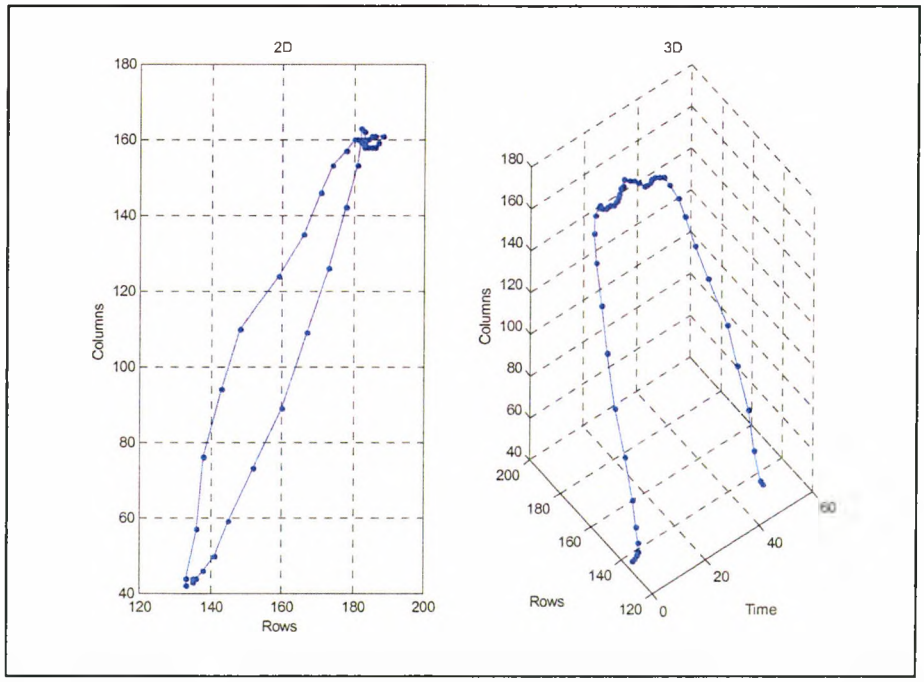


Figure 4.21 2D (left) and 2DT (right) representation of a gesture trajectory

The length of the gesture sequence is normalised using multirate techniques to permit comparison of gesture harmonics by the application of the FFT (Fast Fourier Transform) to the trajectory data. The multirate process is used to directly re-sample the gesture sequence to a fixed number of samples, for all gestures. This technique allows for gestures taken at different sample rates to be compared. It also allows for compensations to be made for the variations in gesture timing between gesturers. The frequency analysis of a single gesture is similar to the spectrum analysis of a finite length signal. The application of the DFT (Discrete Fourier Transform) to the frequency analysis implicitly implies that the input signal is periodic. In practice the DFT is implemented by the FFT as this algorithm is a fast implementation of the DFT when the number of samples is a power of two. It is inappropriate to meaningfully compare harmonics from different sample lengths. The multirate process is used to directly change the sampling rate of the gesture sequence to a fixed number of samples for all gestures. This enables the harmonics for each gesture to be compared on a like for like basis. The target length of a gesture, N was set at 64, and allows a typical gesture of approximately two seconds to be recorded at 25 to 30 frames per second.

4.8.1. Decimation and Interpolation

The primary multirate processing operations are decimation and interpolation and described in detail by Ifeachor and Jervis (1993). Decimation reduces the sampling rate by an integer factor 'M' or reduces the sampling rate from f_s to f_s/M. To prevent aliasing at the lower rate a digital filter is required to band-limit the input signal to less than half of f_s/M. Instead of the default (Matlab) eighth-order, low pass, Chebyshev type filter, a 30-point FIR (Finite Impulse Response) filter was found to be better suited to the application. The filter decimates the input sequence in only one direction. It ensures that the new samples coincided with the start position of the original samples, and thus do not introduce additional phase shifts to the resulting harmonic analysis.

Interpolation increases the sampling rate by an integer factor 'L' to Lf_s. The signal is low-pass filtered to remove image frequencies created by the rate increase. It is always advisable to have the interpolation stage before the decimation stage, as decimation may remove some of the desired frequency components.

4.8.2. Ratio calculation

Cascading an L rate interpolator with an M rate decimator can make possible non-integer values of sample rate changes. It is also convenient to reduce the overall decimation or interpolation factor into the product of smaller factors. Interpolation and decimation values were restricted to a maximum of 13 as there is a possibility of instability in the low-pass filters with higher orders (Matlab Signal Processing Toolbox). It was found that many ratios could easily be obtained with just one ratio value of L/M. A greater coverage with less error was found using two cascaded ratios of L/M i.e. L_1/M_1 and L_2/M_2.

The target ratio value for the L/M value is obtained by dividing 64 (the normalised gesture length) by the number of samples in the gesture. For example, if the sample

length was 48, the length divided into 64 gives a ratio of 4/3, which means that normalisation occurs by an interpolation factor of 4 and a decimation factor of 3. However, many gesture lengths when divided into 64 do not give a exact values from two integers and so the integer values for 'L' and 'M' have to be selected from a range of values that produce a value within a tolerance of the target ratio value. Possible ratios were calculated finding all ratios equal or greater than one with just the numbers from 1 to 13. For instance 13 could be divided by 13, 12, 11 etc. and 12 could be divided by 12, 11, 10 etc. to give a range of values Some values would occur more than once and so the smallest integer values would be selected e.g. 6 divided by 5 would be chosen instead of 12 divided by 10. In all there were some forty different values produced from the combination of the thirteen numbers. In order to achieve more accurate results the L/M ratios were multiplied by every possible combination of themselves to produce a matrix of a larger number of possible ratios. This is implemented by cascading the two L/M ratios together to give the overall ratio of L_1/M_1 x L_2/M_2.

No.	Sample length	L_1	M_1	L_2	M_2.	Overshoot
1	28	2	1	8	7	0
2	30	4	3	8	5	0
3	31	9	8	11	6	1
4	47	7	6	7	6	1
5	53	9	8	13	12	1
6	58	13	12	1	1	-1
7	64	1	1	1	1	0

Table 4.1 Example of Interpolation (L_1, L_2) and Decimation (M_1, M_2) factors required to normalize sample lengths to 64.

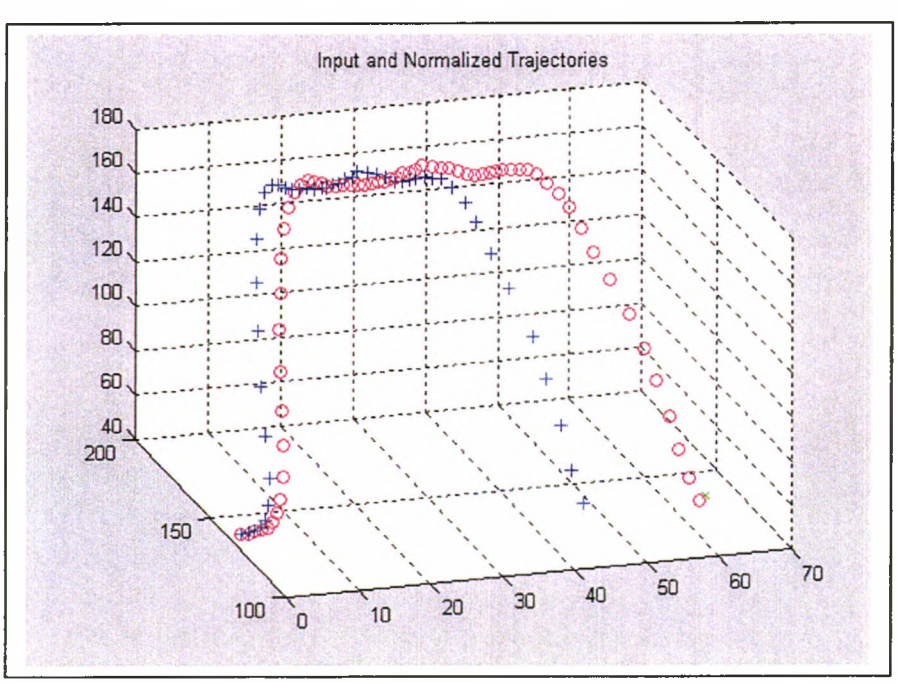


Figure 4.22 2DT view of the normalization of a 47 sample (blue cross) gesture to 64 samples (red circle) by the ratios given for No. 4 of Table 4.1

A selection of ratio values to a tolerance of 0.01% is shown in Table 4.1. However, in practise cascading the two interpolator–decimator pair normally results in either the number of samples being on the target length of 64 or just one sample over at 65. Ratios were calculated from 11 samples to 66 samples. The tolerance could be maintained until sample values close to 64 were encountered. If further refinement were needed an additional ratio factor could be added. It was thought unnecessary to do this for this application as only a few ratios were affected. The result of normalization of a 47 sample gesture trajectory is shown in Figure 4.22.

4.8.3. Aliasing considerations as a result of normalisation

Increasing the typical gesture from, for example, thirty samples to sixty-four samples will not change the frequency resolution of the gesture. However, with sixty four samples it is established that there will be information about the gesture from thirty two harmonics. It would seem that there is more harmonic information now available than from, for example, a gesture of length of thirty samples that would generate fifteen harmonics. It is found that most of the high frequency data is ignored as the gesture is mainly characterised by the low frequency content. The overall frequency content is governed by the frame rate that is normally twenty-five frames per second. This means that the first twelve harmonics are those that can be reliably recorded for authentic characterisation of the gesture. The only possible cause of aliasing is when considering gesturers that are longer in duration than expected with superimposed oscillations (as described in Chapter 7) and sub-sampling is required to accommodate the normalisation process based on a length of sixty four. The generation and characteristics of harmonics generated from the normalised data is discussed in the next chapter.

4.9. Summary

The general hypothesis of that there being just three main skin-coloured moving areas in a sequence is confirmed by the experimentation, when there is just one gesturer in the scene. In complex environments with good lighting the data generation techniques worked well. The techniques with SCMI and SCMIE objects, for the majority of the sequence, generate one object that represent the mid-point of a silhouette of the hand defined by the skin-colour mask.

When lighting conditions are poor, the techniques incorporating SCMI, SCME and SCMIE objects tended to fail. Data is not generated for many of the frames because of either poor segmentation or because of threshold value becoming invalid. However, the basic process, that generates the SCM objects, continued to function. The rank ordering of the data maintains a high expectancy that either the most significant object or the second most significant object will locate the position of the relevant gesturing motion. The cost of using the ANDing output is the fragmentation of the gesture objects, especially in the hand region. Tests carried out in these poor conditions showed that using 'opening' or 'hole-filling' did not improve data. It became apparent that neither had any major benefit. When there was only small amounts of movement, the 'opening' removed noise-like objects and so resulted with less data. In general the optimal ranges of both Hue and Saturation gave the best set

of data, but were closely followed by using just the Hue range with the default Saturation setting.

In these initial experiments the pose of the gesturer, although not controlled, was some what static and most of the movement in the sequence came from the dominant hand. In more realistic environments (PETS sequence) there was much more movement from the gesturer and the two other people in the same image sequence. It became apparent that the original hypothesis of three skin-coloured and motion objects in an image could be extended and the OSA was more versatile than originally expected. The tracking of SCM objects was tested in this more complex scene and the ability to track one person gesturing was maintained, even under long sequences when one of the other people's movements was more pronounced and was being indicated as the most significant object. For the majority of the time the OSA worked effectively and could substitute for the second object instead of the first object where appropriate. At this stage there does not seem to be the need to incorporate the third object in the tracking process, as in the case of little movement the substitution of the previous sample data is adequate. However, it could be incorporated into the algorithm if required. In addition it was found possible to track two hands simultaneously, but only if the hands were kept apart. However, as the main emphasis of this thesis concerned one-handed gesturing this avenue was not explored further at this stage.

The overall aim of tracking technique was to produce a simple, effective hand tracker that would generate data sufficiently robust to enable subsequent trajectory shape analysis. The use of the OSA enabled a range of hand gestures to be tracked in a number of different environments and with different gesturers that was simpler to implement than other popular tracking mechanisms.

The technique for normalising the length of a gesture sequence to 64 samples worked over a range of sample lengths using a pair of interpolation and decimation ratios. There were a number of gesture lengths where it was not possible to change the length to exactly 64 and often it was 65. A full analysis of how this overshoot affected the data in the frequency domain is discussed in the next chapter. In addition, in the next chapter the truncations of the data and issues with the segmenting of the gesture trajectories are analysed.

5. Fourier Analysis of Gesture Trajectory

This chapter focuses on the frequency analysis of a gesture modelled as an aperiodic waveform. 1D and 2D frequency analysis is explained and then developed for the 2DT case. In 1D analysis the concept of frequency components constructed from exponential equations giving rise to positive and negative sequence components is made and visualised in the time domain as a helical construct. The matrix equations that are used to explain the 2D recognition of objects by equally spaced samples in the spatial domain (Fourier Descriptor technique) are modified for equally spaced samples in the time domain. Exponential equations are developed to show the positive and negative sequence components that represent 2DT motion. The subsequent analysis produces elliptical structures in the appearance-based view or 'elliptical corkscrew' in the time domain. In the event that the visualisation of these structures is difficult to understand, many examples are shown to explain the process. The appropriate normalisation of the frequency coefficients allows for scale and translation invariance, allowing for automatic adjustments to take place from sequence to sequence. Analysis of the frequency components shows that phase is an important parameter that has components from the spatial domain and the time domain. Importantly each harmonics is found to have a unique 'orientation' angle in the spatial domain. The characterisation of many different types of gesture trajectory is investigated and explained in terms of the properties of the harmonic components. The synthesis of a gesture trajectory made from the first three harmonic components is shown to closely match the original trajectory path. Finally the 'orientation angle' properties are used to explain and to recognise a series of five 'pointing' gestures made by six different gesturers.

5.1. Fourier Analysis Applications

The application of Fourier analysis to a range of problems is not new. Fourier's seminal work on heat flow problems using trigonometrical series was first published in 1807 (Lynn, 1994). Fourier analysis has been applied extensively to 1D problem such as spectrum analysis, convolution and data communications. The identification of 2D shapes, known as the Fourier Descriptor technique (Zahn and R. Z. Roskies, 1972, Kuhl and Giardina, 1982) uses in complex form, the coordinates of equally spaced points around the perimeter as input to the FFT, the input data being in the spatial domain. The resulting harmonics are given in complex form that includes phase information relating to the rotational position of the sample starting point. Simple manipulation of the harmonic data makes it scale and translation invariant by scaling all harmonic amplitudes by the first harmonic component and removing the d.c. term, respectively. A shape of any size or position can be represented by the magnitudes of a number of harmonics. Harmonic profiles of different shapes can be compared against standard shapes to classify the shape of the unknown object. The Fourier Descriptor is thus a useful tool in the recognition of a shape of a closed planar figure.

Fourier Descriptor techniques have also been applied to gait recognition (Mowbray and Nixon, 2003). The Fourier descriptors model the boundary of a silhouette, using a fixed number of samples for every frame. The gait signature is calculated from the

spatial-temporal Fourier descriptors for a sequence. The majority of the information about the subject's gait is contained in the low-order descriptors. Masters (1994) describes an object recognition system using frequency domain data for input to neural network and found that just 12 harmonics from the 256 samples were adequate for representing a demanding 'T' shape. Chen et al. (2003) use a hand tracking system, similar to the work of this thesis. The hand gesture recognition system recognises dynamic gesture performed singularly in a complex background using the fusing of skin-colour, motion and edge cues. From hand tracking, shape information of the hand by Fourier Description techniques and motion features are input to a HMM for the recognition process. Wallace and Mitchell (1980) showed how a 3D object (aircraft) could be represented by a library of 2D normalized projections. An interpolation procedure in the frequency domain gave more accurate determination of the angle of the object than by simply taking the orientation of the nearest library projection. Tracking of the aircraft trajectory is accomplished by identifying the aircraft in each frame, re-identifying the object and its orientation.

However, Lin and Hwang (1987) show that an alternative representation of the Fourier series is the elliptic Fourier features that may be expressed in matrix form. Lin and Hwang (1987), explain the full mathematical implication of the representation of shape from a set of ellipses. The centre of the u+1st ellipse revolves around the u^{th} ellipse. The revolving frequency of the u^{th} ellipse is u times the first ellipse. The locus of the last ellipse is the contour of the shape. It should be noted that the orientation angle of the u^{th} ellipse is the sum of all the preceding orientation angles. Lin and Jungthirapanich (1990) take the application of elliptic Fourier series a step further and apply it to 3D object recognition. The resulting invariants are implicit functions of the major and minor axis as well as the angles defining the relative orientations. These invariants can be used as features for object recognition. These invariants performed well using simulated and real images. Results also showed that the invariants were insensitive to random noise, because the high frequency noise is dropped when the Fourier series is truncated.

This chapter describes how 2DT (Two dimensions and Time) Fourier analysis technique is applied to gesture trajectories which are aperiodic in nature. The analysis of trajectories, normalised to the same length, allows harmonic content to be readily compared. The harmonic content is considered in a number of ways, from its complex form to the magnitude and phase interpretation. It is found that the trajectory or waveform can be described as an infinite series of harmonic components at different orientation angles. These harmonics can be visualised as two-dimensional ellipses in the x-y plane or as 'elliptic corkscrews' in 2DT space. The variation of the attributes of the harmonics can synthesis many trajectory shapes. Interestingly, the analysis of the properties and characteristics of the 'elliptic corkscrews' in a range of possible gesture trajectories is similar to the features established by Gibet et al. (2001). The studying of some 1359 signs of the FSL (French Sign Language) database showed that for 80% of the Signs, the configuration is static i.e. the fingers do not move when other parameters vary. In addition many of the configurations could be described from a smaller sub-set. However, movements were categorized into 5 main primitives as shown in Table 5.1

Primitive	Proportion (%)	Primitive Description
Line	42.9	The hand's trajectory in space is a straight line
Arc	26.1	The hand follows an arc in space
Static	17.3	The arm is motionless during the gesture
Circle	10.9	The hand's trajectory is an ellipse
Complex	2.8	The trajectory is more complex than in the other primitives, or is composed of several primitives (movements in zigzag, waves, spirals, etc)

Table 5.1 Proportion of different movement primitives (Source, Gibet 2001)

The work described in this thesis takes the actual hand location data as input to a recognition system in which the time domain data is time-normalised by multi-rate methods to a constant number of samples before being transformed into the frequency domain. Gesture action is viewed differently to that of the state based perspective. It is recognised that gestures are an aperiodic action having near identical starting and stopping locations. A view of gesture action is similar to tracing around the perimeter of an object once and is hence very similar to Fourier Descriptor methods. However, in the 2DT application, sampling takes place of coordinates in the time domain and so the trajectory coordinates are not constrained to a closed planar surface as required by spatial Fourier Descriptor techniques. In this application, the rotation of the ellipses are observed over time and visualized as 'elliptic corkscrews', i.e. the u^{th} ellipse having 'u' revolutions uncoiled. A trajectory is not necessarily a planar motion and hence 'elliptical corkscrews' will be oriented at different angles to each other. The properties that determine the constructed contour of each 'elliptical corkscrew' is the orientation angle given by the major axis of the ellipse; the relative values of the major and minor ellipse lengths and the starting phase of each ellipse.

5.2. Fourier Analysis Concepts

The transformation of aperiodic time domain data to the frequency domain is well established by the Fourier Transform for continuous signals and the DFT (Discrete Fourier Transform) for discrete signals. The DFT is invariably implemented by a FFT (Fast Fourier Transform) to gain the advantages of a very fast transformation into the frequency domain. Time domain signals that are sampled 'N' times give rise to 'N' harmonics, although for real input data only N/2 harmonic are realised because of the symmetry of the magnitude of the harmonic components. A useful property of the DFT is that the reverse process, the IDFT (Inverse Discrete Fourier Transform) can be implemented to retrieve time domain data from the harmonic or frequency components. With most waveforms, the magnitude of the harmonics diminishes as the harmonics increase and it is found that in the reconstruction process, from frequency domain to time domain, only a few harmonics are required to give a good representation of the original waveform.

The use of the DFT to investigate the frequency components of signals and the frequency performance of systems is widely described and has had a profound influence on many branches of engineering and applied science. Most of these examples relate to 1D application that has varying amplitude with time. Fourier theory was adapted to the 2D problem of shape recognition. In this case the input data represents coordinate data of equally spaced samples around the perimeter of a closed planar shape or object, but represented in complex format. The analysis of the frequency components (Kuhl et al. 1982 and Lin et al.1990) show that each harmonic can be represented by an elliptical structure in the spatial domain. The overall shape is based upon the last point of the highest frequency ellipse where each ellipse is at a fixed orientation with the centre of the $k^{th}+1$ ellipse revolving around the k^{th} ellipse k+1 times. The other most important aspect of Fourier Descriptor analysis is that through suitable simple mathematical manipulation the coefficients of the harmonics can be adjusted for scale and translation invariance so object size, position and orientation is not important when recording data.

This chapter explores gesture trajectories as 2DT (Two Dimensions and Time) aperiodic signals. The representative coordinates of the hand are described by complex numbers in the spatial domain but sampled at regular intervals in the orthogonal time domain. The resulting frequency analysis of 2DT signal can be interpreted using knowledge gained by the analysis of frequency components in both 1D and 2D applications. The analysis and understanding of 2DT signals is gained through using the concept of positive and negative frequency sequence components as well as through the more traditional trigonometrical matrix equations. The Fourier analysis of the 2DT data in its complex number form gives minimal information as to the characteristic of the gesture, except that generally as the harmonics increase the amplitudes decrease. Interpretation of the harmonic data after suitable magnitude normalisation and phase manipulation of the harmonic data gives insight into the shape and nature of the gesture trajectory.

Following the analysis of a gesture trajectory a synthesis of the trajectory can be constructed by adding a series of 'elliptical corkscrew' structures representing each harmonic component. Each 'elliptical corkscrew' is at a given orientation angle. The number of rotations is equivalent to the harmonic value.

For this thesis, interpretation of the harmonic data was investigated in a number of ways involving standard signal shapes, simulated trajectories and actual gesture trajectories. Simulated gestures formed ideal straight line, arced and oscillatory trajectories. One of the interesting results of this investigation was the role of phase in characterising a trajectory and the orientation angle that reflected the harmonics position in the spatial domain.

The formation of ellipse characteristics is considered using positive and negative sequence components. The magnitude of the major and minor axis of the ellipse is related to the relative amplitudes of the positive and negative sequence components. The relative amplitude of the positive and negative sequence components also signify the direction of rotation and the harmonic value determines the number of revolutions. In the time domain the rotation of the ellipse can be visualised as an 'elliptical corkscrew'.

An investigation into the phase information shows that it is made of two components; one that relates to time domain discontinuities and the other component that relates to the 'orientation' angle. The lower order harmonic components are found to characterise the shape of the gesture trajectory. Virtually all of the signing primitive description (Gibet, 2001) can be confirmed by the interpretation of the harmonic amplitude and phase components of the ellipses.

The harmonics are also used to assess imperfections in the trajectory. Gesture trajectories that do not return to the starting position produce a discontinuity in the waveform and cause a truncation error showing up as spectral leakage and phase shift in the time domain. Additionally the object selection algorithm can make mistakes. This means a limited tracking error can be tolerated as it appears as a high frequency signal which can be disregarded in gesture characterisation.

5.3. Fourier Analysis in 1D, 2D and 2DT domains.

5.3.1. One-Dimensional (1D) Fourier Analysis

The Discrete Fourier Transform (DFT) and its inverse (IDFT) are stated: -

$$DFT,\ X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi}{N}nk} \qquad n = 0,1,\ldots N-1$$

$$IDFT,\ x(n) = \frac{1}{N} \sum_{k=0}^{N-1} X(k) e^{+j\frac{2\pi}{N}nk} \qquad k = 0,1,\ldots N-1$$

where k is the number of harmonics, n the data points and N the total number of samples.

The numbers in table A5.1 of Appendix V illustrate the method of calculation of the harmonic values X(k) for the input data x(n) of values [1, 0, 0, 1]. It is worthy to note that X(0) represents the sum of the input data values and X(1), X(2) and X(3) represent the harmonic values. In this calculation X(2) is equal to zero and X(1) and X(3) are equal to 1+j and 1-j respectively. The interesting thing about the first and third harmonics are that they form a complex conjugate pair where the amplitudes are the same but the phase are equal but of opposite polarity, i.e. X(1) has an amplitude of $\sqrt{2}$ and a phase angle $\varphi(1)=+45°$ and X(3)) has an amplitude of $\sqrt{2}$ and a phase angle $\varphi(1)=-45°$. The fundamental angular frequency is calculated by reference to the sampling interval, T where $\omega = 2\pi / T$.

Euler's formula shows how the complex formula can be expressed in polar form: -

$$re^{j\theta} = r(\cos\theta + j\sin\theta)$$

Alternatively the argument θ can be expressed in terms of angular frequency ω and time t so $\theta = \omega t$: -

$$re^{j\omega t} = r(\cos\omega t + j\sin\omega t)$$

Additionally, the negative form is written as: -

$$re^{-j\omega t} = r(\cos\omega t - j\sin\omega t)$$

It is also useful to consider cosine and sine expressions in their exponential form as:-

$$\cos(\omega t) = \frac{1}{2}\exp(j\omega t) + \frac{1}{2}\exp(-j\omega t)$$

$$\sin(\omega t) = \frac{1}{2j}\exp(j\omega t) - \frac{1}{2j}\exp(-j\omega t)$$

These equations illustrates that a cosine signal of unit amplitude and angular frequency ω can also be represented by double-sided spectrum of a positive and negative frequency component of amplitude of half the single sided representation, as shown in Figure 5.1. Phase shift can also be included in the notation so that cosine waveforms with phase shift of ϕ results in the following equation: -

$$\cos(\omega t + \phi) = \frac{1}{2}\exp(j\omega t)\exp(j\phi) + \frac{1}{2}\exp(-j\omega t)\exp(-j\phi)$$

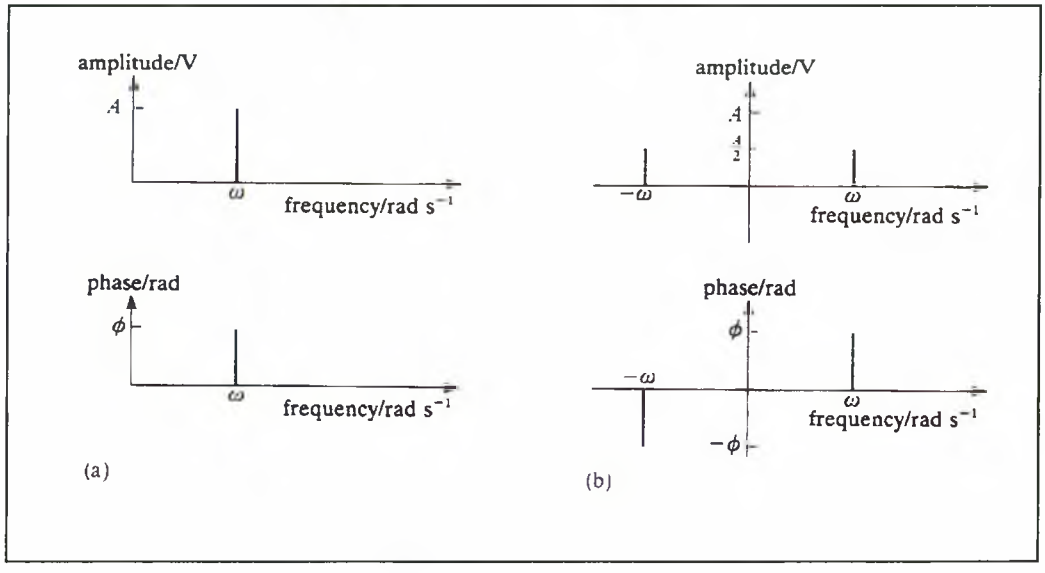


Figure 5.1 Single- and double-sided spectra of cos(ωt+ϕ). (Source: Bissell and Chapman, 1995)

The previous equation shows that the harmonic components can be represented in exponential form. The value of X(1) with a value of 1+j can be represented by an exponential component representing the angular frequency and an additional exponential component representing the phase shift of 45° or π/4. In the example described in Table A5.1 the second harmonic term (1+j) is said to be rotating with a positive frequency and the fourth harmonic term (1-j) is said to be rotating with a negative frequency. Figure 5.2 shows a complex exponential, $f(t) = e^{-j\omega t}$, negative frequency, as a phasor diagram and time domain representations (Kraniauskas 1993, Transforms in Signals and Systems, Addison-Wesley).

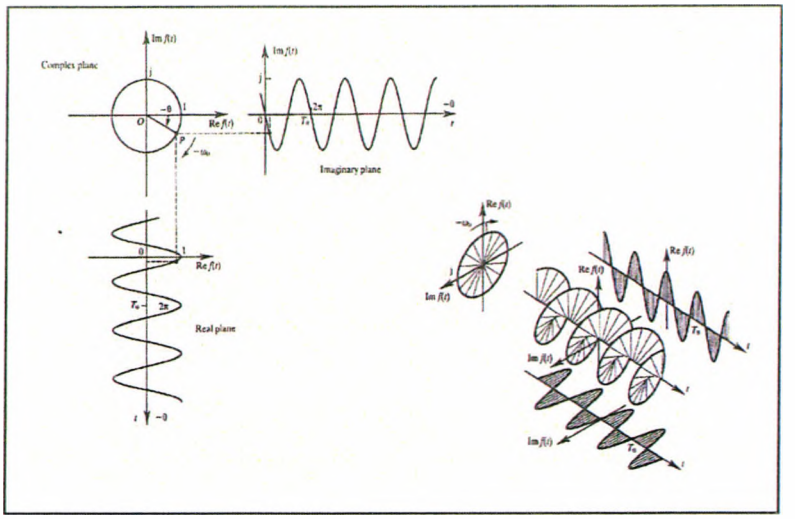


Figure 5.2 Complex exponential, $f(t) = e^{-j\omega t}$, negative frequency (Source: Kraniauskas,1993)

An interesting feature of Kraniauskas's (1993) treatment is the 2DT representation of the rotating phasor over time as a helical structure. This structure is very similar to the more general 'elliptical corkscrew' structure which is developed in later sections of this chapter that aids in the analysis of gesture trajectory data.

5.3.2. 2D Fourier Analysis – Fourier Descriptor

Lin and Hwang (1987) explain that Fourier descriptors are very useful for describing the shapes of closed contours. The periodic function which is obtained by tracing the closed contour can be expressed in a Fourier series. A curve is represented as a function of arc length by the accumulated change in direction of the curve. The function is expanded in a Fourier series for recognition. The closed contour is represented by the function of the arc length as $[X(l), Y(l)]$ and represented by the complex periodic function $X(l), iY(l)$ which is expanded in a Fourier series. A similar representation called the elliptic Fourier features (Kuhl and Giardina, 1982) expands $X(l)$ and $Y(l)$ separately and puts them into matrix form. Consequently a shape is interpreted as a specific composition of ellipses. Lin et al.'s mathematical treatment is detailed in Appendix V.

5.3.3. Gesture trajectory analysis (2DT)

At first sight the view of a gesture trajectory in 2D is very similar to object identification using Fourier Descriptors. However, there are some important differences. Perhaps the most obvious point is that Fourier Descriptors approximate equisampling of space dimensions (2D), whereas gesture trajectories are time sampled. The 2D and 2DT views of a typical gesture trajectory are shown in Figure 4.21 (Chapter 4). A gesture trajectory, when viewed in appearance-based space can appear like the outline or perimeter of an object and hence Fourier Descriptor techniques might be applied to the trajectory to analyse its frequency components.

However, there are important differences between an object perimeter outline and a gesture trajectory. The perimeter of an object in appearance-based space is for a closed planar space and the perimeter does not cross-over any point of the contour as seen in the figure. The actual movement of the gesture is in 3D space and whilst the camera image captures 2D information there is no constraint as to the movement of the hand crossing the path of a previous trajectory track.

The 2DT picture of Figure 4.21 actually shows that the time domain view of the gesture trajectory is very similar to an aperiodic pulse waveform The Fourier analysis of a simple rectangular pulse is well documented in the literature. A pulse of amplitude A and width τ in the limits of $-\tau/2$ and $+\tau/2$, will give a continuous spectrum: -

$$G(f) = A\tau(\frac{\sin \pi f \tau}{\pi f \tau})$$

The Fourier transform G(f) of the pulse g(t) is a unique frequency-domain representation of the pulse and in general takes on a complex value for each value of frequency. Frequency spectrum of pulses often used in communications applications illustrates the spectrum as positive and negative components and because the input data is real the positive and negative spectrum components are symmetrical.

In practice the Fourier components are obtained by digital computation rather than by analogue processing. The transition from a continuous spectrum to a discrete spectrum is explained in most signal processing text books. The resulting DFT (Discrete Fourier Transform) can be used to analyse discrete valued pulses as was explained in the previous section of this chapter. The significant difference in the analysis of gesture trajectories is that the input data is complex. The analysis is concerned with interpreting harmonic content in the spatial and time domains. For example in Figure 5.3 three points A, B, and C lie on an ellipse as seen in the spatial domain but are sampled at regular time intervals t1, t2 and t3 in the time domain. It is interesting to note that the points A, B and C in the spatial domain form an elliptical shape, but only sinusoidal amplitude changes would be seen in the time domain.

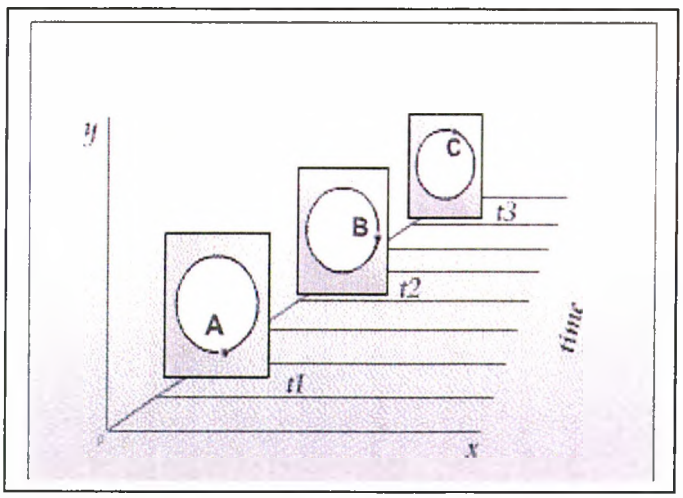


Figure 5.3 Three points A, B and C on an elliptical trajectory in appearance-space sampled at t_1, t_2 and t_3 in the time domain

Further clarification of the concept can be made by the study of the ellipse (blue, top left picture of Figure 5.4). It is interesting to note that in the top-right picture shows the change of 'y' with time as one oscillation of a sine wave; the bottom-left picture shows the change of 'x' with time as one oscillation of a cosine wave and the bottom right shows a 2DT view as one oscillation of an 'elliptical corkscrew' structure. Further clarification of the concept is shown in Appendix V, Figure A5.5 which represents the sampling a waveform produced by the summation of two ellipses at two different frequencies. Visualisation of the 2DT picture appears different depending on the viewing angle as shown in Figure A5.6 and does not capture a consistent shape or interpretation.

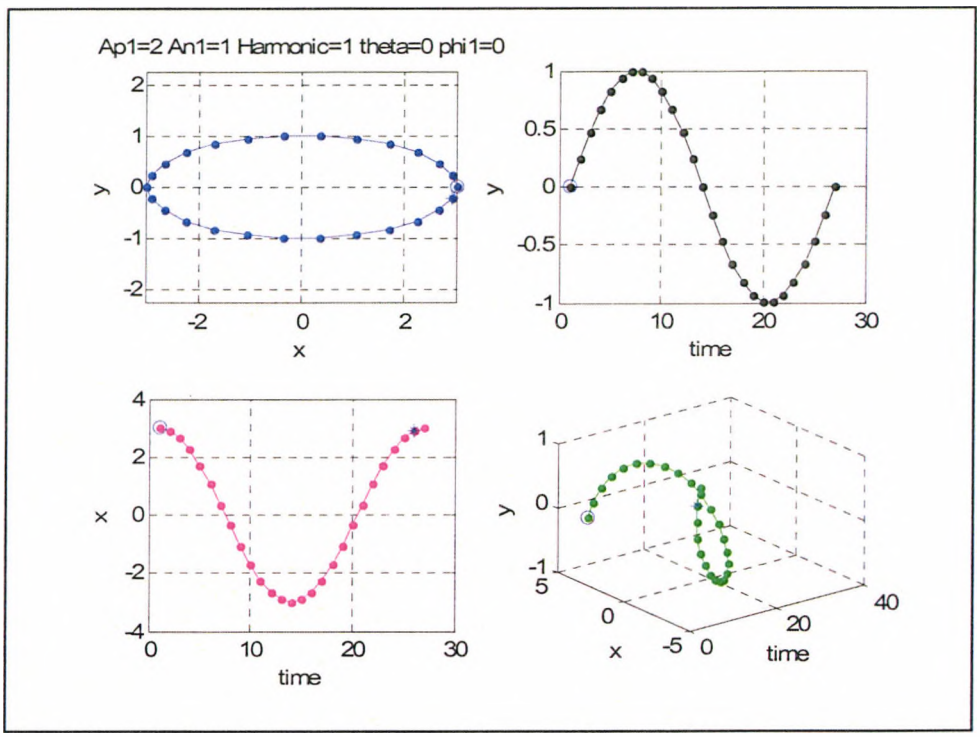


Figure 5.4 Pictures of 4 views of an ellipse sampled in the time domain. The top left picture shows the spatial domain representation; the top-right shows the change of 'y' with time (a sine wave); the bottom-left picture shows the change of 'x' with time (a cosine wave)and the bottom right shows one revolution of the 'elliptical corkscrew', i.e. a 2DT view.

The pictures in Figure 5.4 and Figures A5.5 indicate that the complex data generate two distinct sets of frequency components, one relating to the time domain and the other to the spatial domain. The following mathematical and experimental investigations confirm these observations.

5.3.4. Developing exponential equations (positive and negative sequence components) to describe gesture trajectories in 2DT space

The mathematical equations describing spatial domain realisation of ellipses are stated in Appendix V, where the variable 't' was related to the spatial dimensions. Consideration is now given to the variable 't' being related to time so that instead of tracing points around the perimeter of an object in appearance-based space, a point in appearance-based space (x + jy) can be traced at equally intervals of time. The

feature that is tracked is the centroid coordinates of the hand, a single point in appearance-based space.

The matrix equation (Appendix V): -

$$\begin{bmatrix} x_k(t) \\ y_k(t) \end{bmatrix} = \begin{bmatrix} \cos\theta_k & -\sin\theta_k \\ \sin\theta_k & \cos\theta_k \end{bmatrix} \begin{bmatrix} A_k & 0 \\ 0 & B_k \end{bmatrix} \begin{bmatrix} \cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & \cos\varphi_k \end{bmatrix} \begin{bmatrix} \cos kt \\ \sin kt \end{bmatrix}$$

when expanded and appropriate trigonometrical simplifications have taken place give:-

$$x_k(t) = A_k \cos\theta_k \cos(kt + \varphi_k) - B_k \sin\theta_k \sin(kt + \varphi_k)$$
$$y_k(t) = A_k \sin\theta_k \cos(kt + \varphi_k) + B_k \cos\theta_k \sin(kt + \varphi_k)$$

Instead of sampling in the spatial domain at regular intervals the variable '*kt*' may be changed so that the variable is taking regular samples in the time domain. Using the normal convention for a sinusoidal waveform with variable ϕ_k, becomes: -

$$x_k(t) = A_k \cos\theta_k \cos(\phi_k + \phi_{k0}) - B_k \sin\theta_k \sin(\phi_k + \phi_{0k})$$
$$y_k(t) = A_k \sin\theta_k \cos(\phi_k + \phi_{k0}) + B_k \cos\theta_k \sin(\phi_k + \phi_{k0})$$

or, rotating at 'ω' radians per second the equations become: -

$$x_k(t) = A_k \cos\theta_k \cos(\omega_k t + \phi_{k0}) - B_k \sin\theta_k \sin(\omega_k t + \phi_{k0})$$
$$y_k(t) = A_k \sin\theta_k \cos(\omega_k t + \phi_{k0} + B_k \cos\theta_k \sin(\omega_k t + \phi_{k0})$$

where for a particular harmonic '*k*', the phase variable is ϕ_k or $\omega_k t$ and the phase shift is ϕ_{k0} and the orientation angle is θ_k.

Previously, when considering 1D data it was discovered that the harmonic generated by Fourier analysis came as complex conjugate pairs and an alternative representation was to model the harmonics as positive and negative sequence components. In this case the amplitudes of these harmonics are equal. If we consider the case of the amplitudes of the pairs of harmonics being different, then we have the case that ellipses are formed. The inclusion of the phase shift, ϕ_{k0} and orientation angle, θ_k to both positive and negative phasor equations give for a particular harmonic and gives an alternative equation: -

$$z_k(t) = Ap_k \exp j(\omega_k t + \phi_{k0} + \theta_k) + An_k \exp - j(\omega_k t + \phi_{k0} - \theta_k)$$

$$x_k(t) = \Re(Ap_k \exp j(\omega_k t + \phi_{k0} + \theta_k) + An_k \exp - j(\omega_k t + \phi_{k0} - \theta_k))$$

$$y_k(t) = \Im(Ap_k \exp j(\omega_k t + \phi_{k0} + \theta_k) + An_k \exp - j(\omega_k t + \phi_{k0} - \theta_k))$$

where the ellipse is described by the positive and negative phasors, with

Ap_k and An_k representing their amplitudes respectively.

It can be shown that the relationship between the major and minor axis of the ellipse formed from the matrix equations and the positive and negative sequence components are: -

$$Ap_k + An_k = A \text{ and } Ap_k - An_k = B$$

A comparison of the two methods is shown in figure 5.5, where the major axis of the ellipse, A is equal to 1.5 and the minor axis, B has a value of 0.5 as shown by a blue 'x'. The phasor representation is shown by a red 'o' and is fully coincident with the 'x' values with positive sequence amplitude of 1 and negative sequence amplitude of 0.5. The orientation angle θ_k is at 30° and orients the ellipse away from the horizontal 'x' axis. The phase shift ϕ_k at 60° just indicates the different starting positions of the positive and sequence components, but does not alter the shape of the overall ellipse.

It is interesting to note that a straight line in appearance-based space is a particular example of an ellipse with $A_1 = 1.0$ and $B_1 = 0$ or $A_p = 1$ and $A_n = 1.0$.

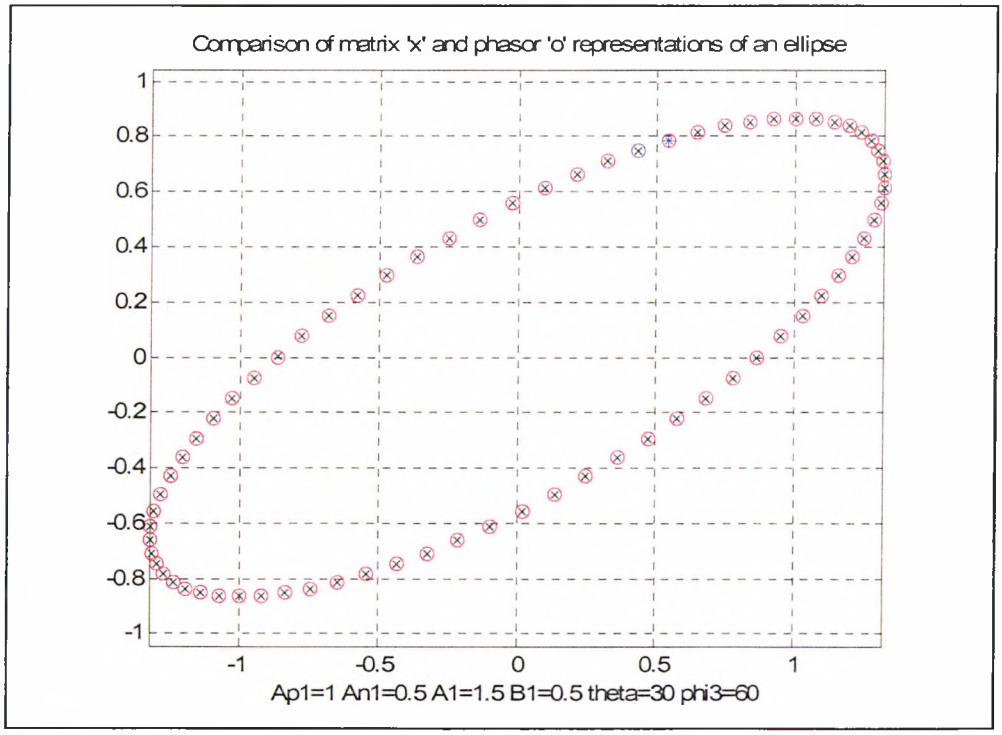


Figure 5.5 Comparing matrix 'x' and exponential 'o' equations of an ellipse with A = 1.5 and B = 0.5 or $A_p = 1$ and $A_n = 0.5$, with Orientation angle of 30° and phase shift of 60°

The effect of the positive and negative components is shown in more detail in the example of figure 5.6. The direction and shape of the ellipse depends upon the magnitude of A_p or A_n. The relative size of A_p and A_n affects the direction of revolution of the ellipses. If A_p is the greater, the ellipse will rotate in an anti-clockwise direction; if A_n is greater, the ellipse will rotate in a clockwise direction. It is noted that the orientation angle θ_k can be found by taking the average of the positive and negative sequence phase. So in general: -

$$\varphi_p = \theta + \phi \text{ and } \varphi_n = \theta - \phi$$

where φ_p and φ_n are the total positive and negative phasor phase shift and θ is the orientation angle and ϕ is the phase shift.

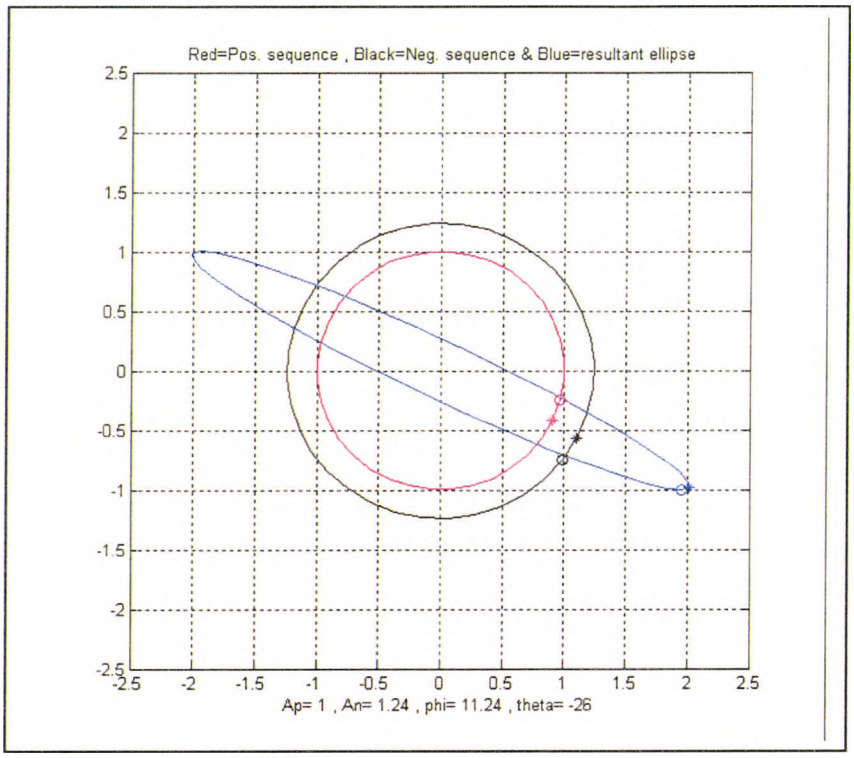


Figure 5.6 Rotating positive (red) and negative (black) sequences and the resulting (blue) ellipse, with 'o' indicating starting point and '*' end point.

This effect of the phase shifts is illustrated in Figure 5.6 where the red circle represents the positive sequence and the negative sequence is represented by the black circle. A symbol 'o' indicates the start of the phasor and the symbol '*' represents the end, so the rotational direction can be ascertained. The resultant ellipse is coloured blue and shows that the ellipse rotates in a clockwise manner in this example where $A_p = 1.00$ and $A_n = 1.24$ and $\phi = 11.24°$and $\theta = -26°$.

The advantage of modelling ellipses with positive and negative sequence components is that the form of the equation is the same as that given by Fourier analysis. Having the same form of equation for analysis and synthesis allows for easy comparisons to be made when modelling trajectories with the results of analysis, as will be seen in later sections.

5.3.5. Exponential synthesis of waveforms

The power of Fourier synthesis is often shown by taking just a few low order harmonics and adding them together to show that a good representation of the original waveform can be made. It can be shown that recombining the first three odd harmonics, as the even order harmonics are zero, from square wave analysis shows

how close the synthesis is to the original square waveform. For the purposes of this investigation it is useful to visualise the waveform both in the 2D (spatial), 1D and 2DT domains. A full explanation of the stages of this reconstruction is shown in Appendix V and in this reconstruction the square waveform only appears in the 'x' plane and in the 'y' plane the output is zero (Figures A5.7, A5.8 and A5.9).

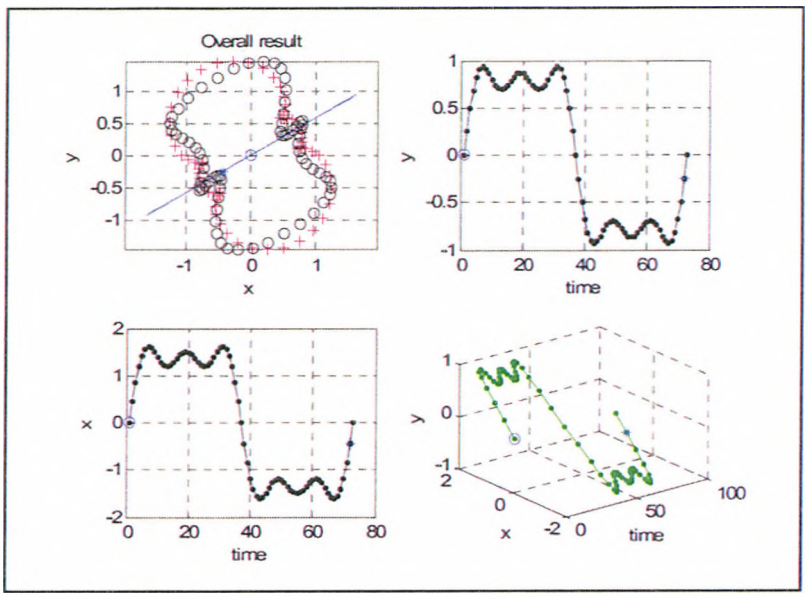


Figure 5.7 Four views of the first, third and fifth harmonics at the same orientation. Rotating positive (red) and negative (black) sequences and the resulting (blue) ellipse. The starting point-and end-point of the time sequence indicated on all pictures as 'o' and '*' (blue) respectively.

The result of adding just the first and third harmonic together but with an orientation angle other than zero is shown in figure 5.7. The 2DT image particularly shows the synthesised trajectory at a slant to the previous figure and also shows a component in the 'y' direction as shown by the waveform appearing in the top-right picture and the blue line in the top-left picture now being at an orientation to the 'x' axis.

The previous examples produced a straight line in appearance-based space (2D), which is just a special case of an ellipse. In the matrix equation the component 'B' is equal to zero or where A_p and A_n have equal values. For illustrative purposes at this stage an example of the positive and negative components being different is shown in Figure 5.8. The amplitudes are at 0.042 and 0.103 respectively and the components are at an orientation of 38° with phase shift of 27°.

Figure 5.8 shows the formation of an ellipse (blue) in the appearance-based picture (top-left). The phase shift is clearly seen in the 'x' vs. time and 'y' vs. time pictures (bottom-left and top-right respectively). The 2DT picture (bottom-right) shows three spirals of a helix type structure described as an 'elliptical corkscrew' and will be referred to later in the analysis of gesture trajectories.

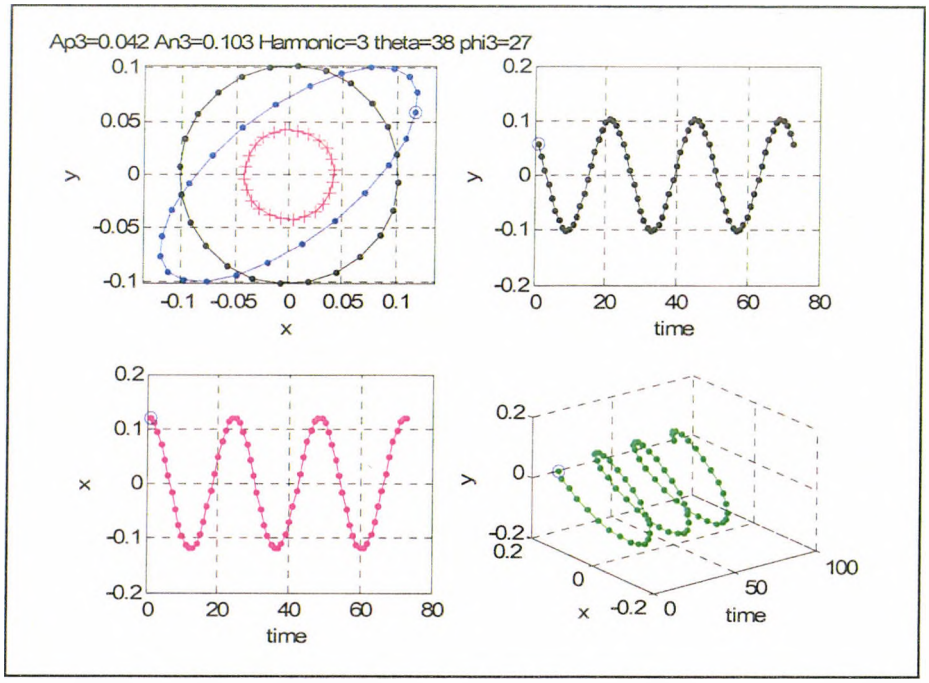


Figure 5.8 Four views of elliptical structure, for a third harmonic, with A_p at 0.042 and A_n at 0.103, orientation angle of 38° and phase shift of 27°.

5.3.6. Cross-over in trajectory contour

Previous discussion had explained that the coordinates of hand gesture trajectories were not the same as tracing around the perimeter of an object because in some circumstances the contour was seen to cross-over. The following discussion explains the nature of the elliptic functions that allows this to occur. The occurrence will be modelled both in the matrix form of the equations and the positive and negative frequency components using just two harmonics.

An example is taken of the first harmonic ellipse having parameters of A1 = 0.5 and B1 = 0.2. The second harmonic ellipse is considered for two conditions A2 = 0.2 and B2 = 0.1 or A2 = 0.1 and B2 = 0.2. It can be seen that for the second harmonic the orientation of the major axis in the first condition the 'x' axis component is greatest and in the second condition it the 'y' axis component is the greatest. The equivalent positive and negative sequence components are as in Table 5.2

Harmonic	A	B	A_p	A_n
1	0.5	0.2	0.35	0.15
2	0.2	0.1	0.15	0.05
1	0.5	0.2	0.35	0.15
2	0.1	0.2	0.15	-0.05

Table 5.2 Comparison of ellipse definition coefficient A and B with A_p and A_n

The picture of Figure 5.9 shows the condition of A1 = 0.5 and B1 = 0.2 and of A2 = 0.2 and B1 = 0.1 (A_{p1}= 0.35, A_{n1}= 0.15; A_{p2} = 0.15 and A_{n2} at 0.05). The top right appearance-based picture shows the overall result of adding the two ellipse structures together as shown by the blue contour formed from the red (positive) and black

(negative) sequence components. The top-left and bottom right pictures show how the 'y' and 'x' components vary with time, respectively. The bottom right picture is a 2DT view of the overall shape.

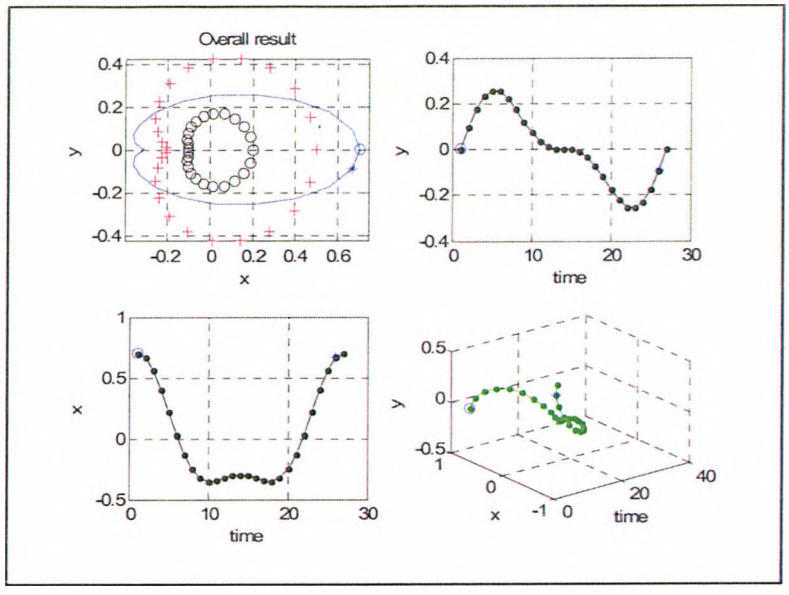


Figure 5.9 Four views of the positive and negative sequence components of the addition of the first and second harmonics (Ap1= 0.35; A_{n1}= 0.15; A_{p2} = 0.15; and A_{n2}=0.05 being equivalent to A1=0.5; B1=0.2; A2=0.2 and B2= 0.1).

The individual first and second harmonic components are shown in Figures 5.10 and 5.11, respectively.

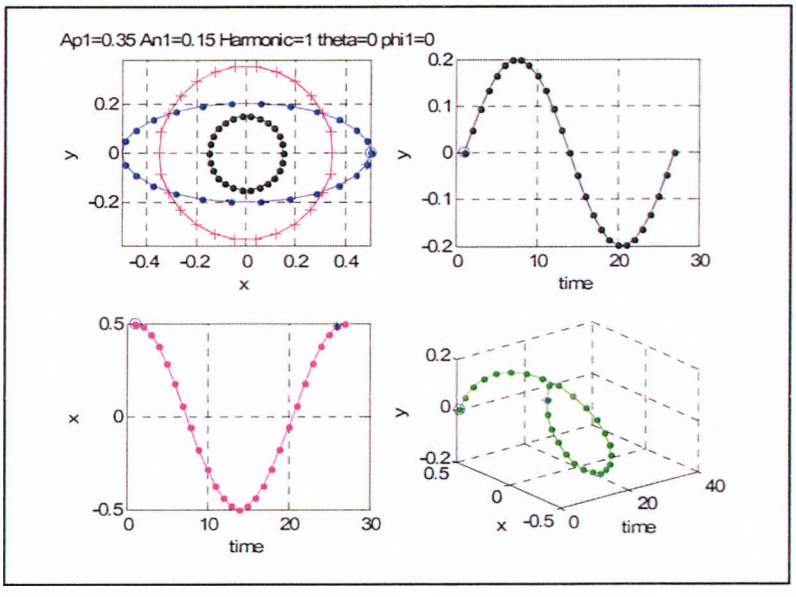


Figure 5.10 Four views of the first harmonic positive and negative sequence components A_{p1}= 0.35 and A_{n1}= 0.15 being equivalent to A1 at 0.5 and B1 at 0.3

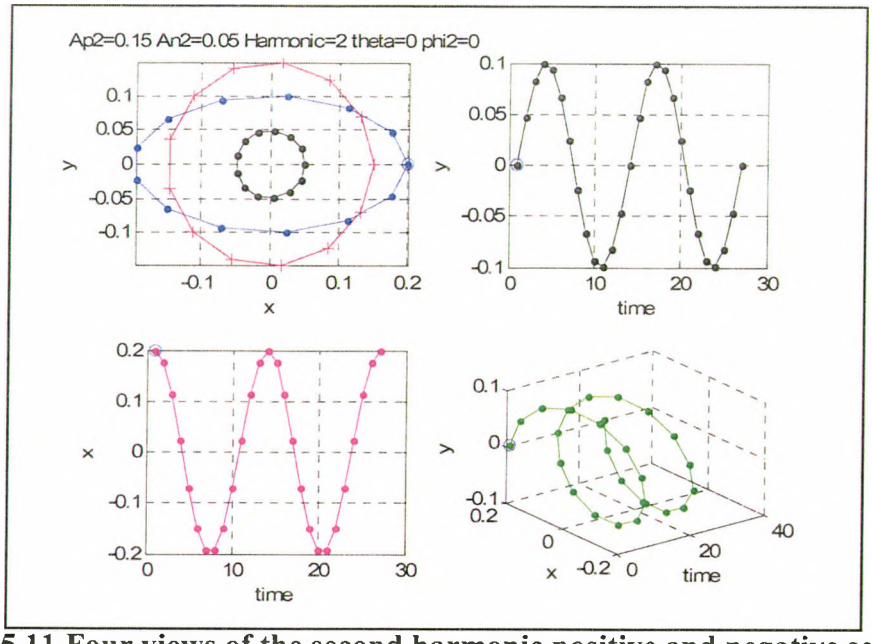


Figure 5.11 Four views of the second harmonic positive and negative sequence components Ap2 at 0.2 and A_{n2} at 0.05 being equivalent to A2 at 0.2 and B2 at 0.1.

It is interesting to note that in both these figures that the ellipse (top-left) lie at the same orientation i.e. the maximum dimension of the ellipse is in the 'x' direction. In addition the 'y' vs. time (top- right picture) shows the sinusoidal waveform and the 'x' vs. time (bottom- left picture) the cosine waveform at the amplitudes corresponding to the A and B amplitudes for that ellipse. The bottom right picture shows the 'elliptical-corkscrew', 2DT representation of each harmonic, the number of revolutions in line with their harmonic value.

When the second harmonic values change to A2 = 0.1 and B2 = 0.2, (A_{p2} = 0.15 and A_{n2} at -0.05) a cross-over in the contour is seen as in the appearance–based view (top-left picture) of Figure 5.12, but not seen in the time domain. The similar result to Figure 5.12 is also obtained using positive and negative sequence components as in Figure 5.13. The second harmonic (Figure 5.14) is now clearly different, to that shown in the previous figure, Figure 5.11. In this case the orientation of the second harmonic is at right-angles to the previous case as the major axis relates to the 'y' axis. In the 'y' vs. time and the 'x' vs. time pictures the amplitudes of the sinusoidal and cosine components have reversed.

To aid clarity of how the time domain feature are perceived over time the 2DT representation is viewed at different angles as in Figure 5.15. All views in Figure 5.15 have a similar elevation views ranging from 0° to -4° whereas the azimuth start at -90° at view 1 and reduce through the views until view 4 the azimuth is at -10°. The pictures indicate the cross-over in 2D space but as the viewing angle changes toward a time domain view the coordinates show the cross-over as an inflection. The interesting characteristic to note is of how ellipses in 2D are seen differently in a time domain view.

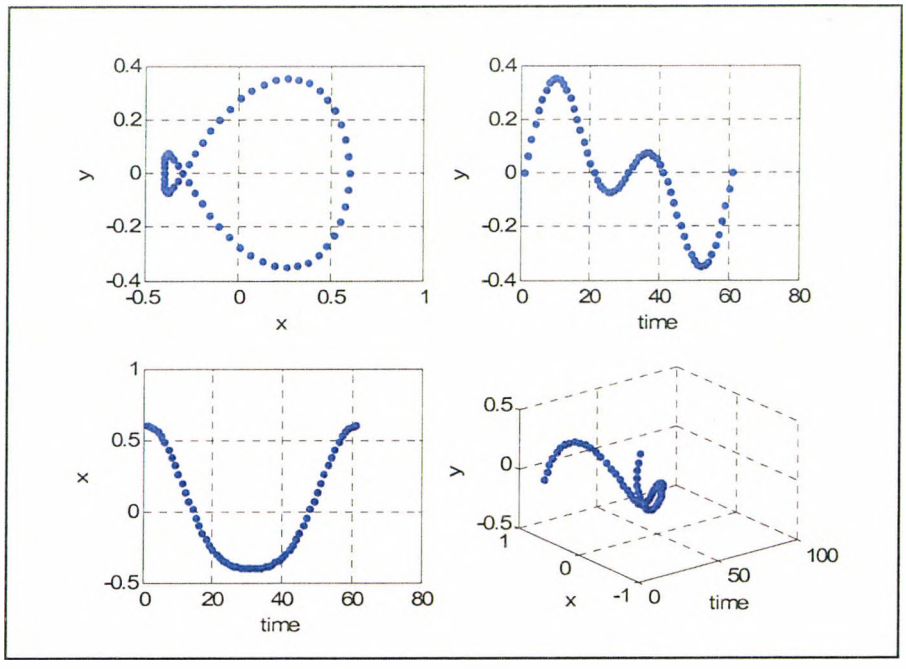


Figure 5.12 Four views resulting from the combination of two ellipses having parameters of A1 at 0.5 and B1 at 0.2 and A2 at 0.1 and B2 at 0.2.

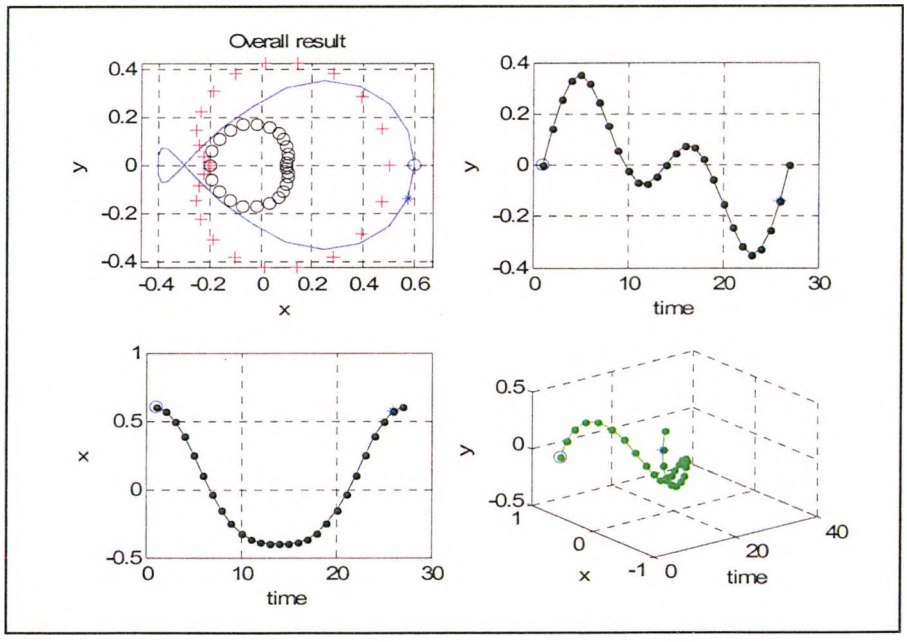


Figure 5.13 Four views of the positive and negative sequence components with A_{p1}= 0.35 and A_{n1}= 0.15 and A_{p2} = 0.15 and A_{n2} at -0.05 being equivalent to A1 at 0.5 and B1 0.2 and A2 at 0.1 and B2 at 0.2.

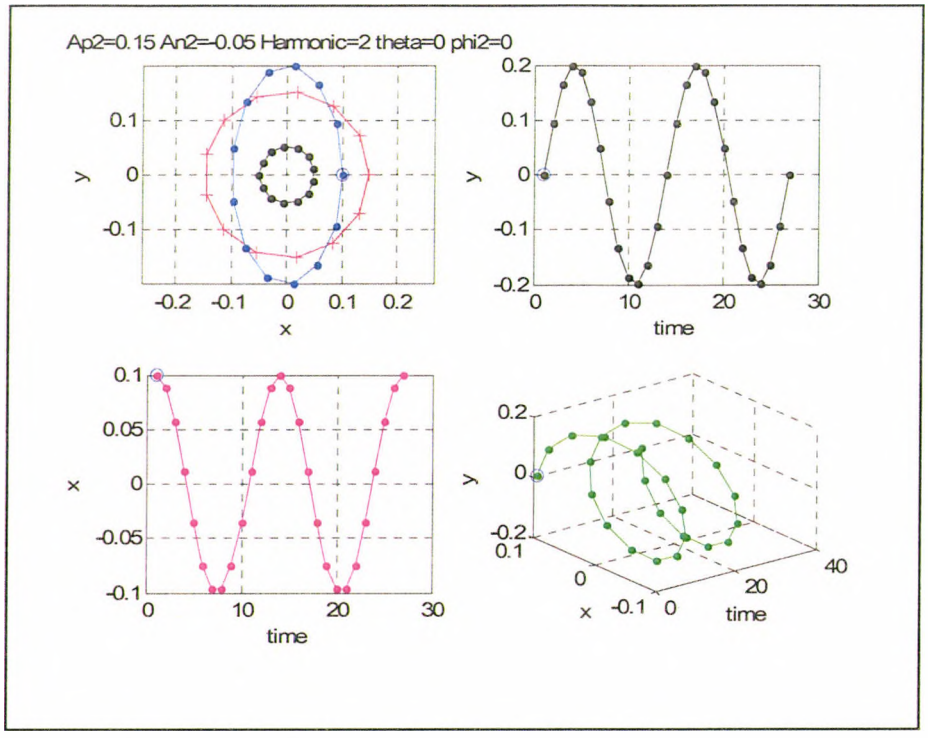


Figure 5.14 Four views of the second harmonic positive and negative sequence components A_{p2} at 0.15 and A_{n2} at -0.05 being equivalent to A2 at 0.1 and B2 at 0.2.

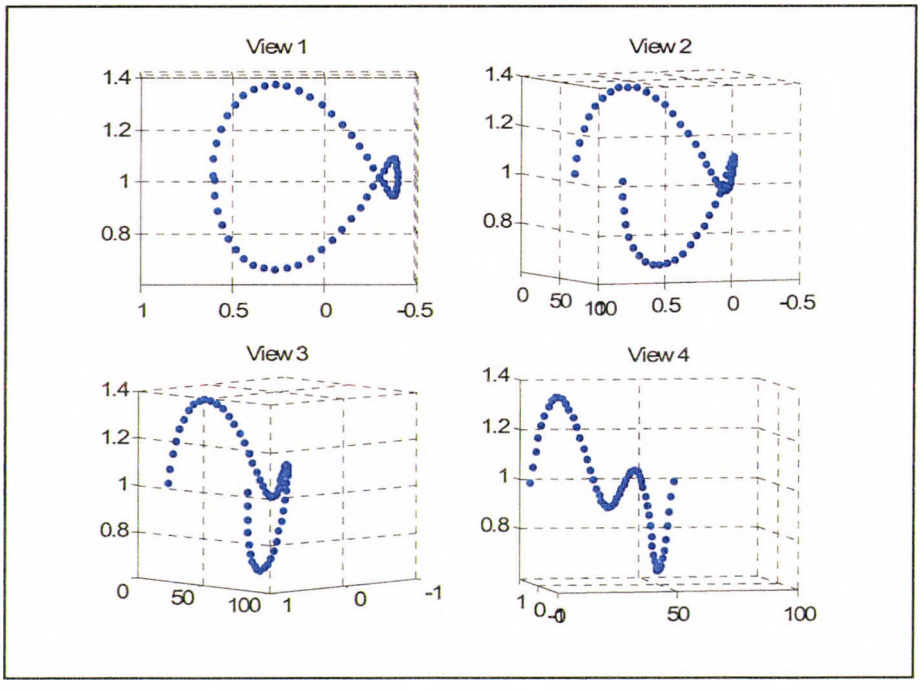


Figure 5.15 Different azimuth views and similar elevation views of two ellipses that produce a cross-over in the contour (Azimuth -View 1 =-90°, View 2 =-69°, View 3 =-48°, View 4 =-10°; Elevation (typical) = -2°)

5.3.7. Translation, Scale and Orientation, θ_k considerations

The previous section showed how gesture trajectories could be constructed from elliptical functions using equations based on positive and negative sequences. A full investigation of the synthesis and analysis has not be addressed at this stage (see section 5.5) as it is worth considering various factors that make the analysis easier to implement and compare.

Firstly, the coefficients of Fourier analysis are complex numbers each having magnitude and phase. To make the gesture independent of position the d.c. term is removed from the analysis. Hence the trajectory is independent of its position in the image.

The analysis is made invariant to the size or recorded values of the trajectory by scaling all harmonic coefficients with respect to the first harmonic, which is usually the largest harmonic. When harmonics generate shallow ellipses it is convenient to select the positive sequence amplitude for the normalisation parameter. In many cases the negative sequence amplitude is very similar in value. However, when the first harmonic shows sign of a pronounced ellipse the largest value of the positive or negative component should be selected for normalisation. The choice of normalisation parameter is discussed further in chapter 7 when other criteria are discovered with more unusual gesture trajectories.

The starting point of a gesture is always at the same point with reference to the resting condition for the stationary hand. This is in contrast to object recognition methods where the starting point can be at any point on the perimeter of the object and so some kind of phase normalisation process is required if phase information is required for analysis. For example, Masters (1994) explains that this can be a difficult and somewhat arbitrary process. The two conditions of starting point phase and rotational phase normalisation have to be simultaneously met. Wallace and Wintz (1980) also discuss an overview of phase normalisation. Phase normalisation has not been found necessary in this research because of the invariance of the starting position.

However, the phase can give important information especially with regard to the orientation of the gesture. θ_k, in gesture recognition application is referred to as the orientation angle as this is the angle that the harmonic makes in the spatial, appearance-based domain. It is usual to have the first harmonic dominate the magnitude of the other harmonics and hence the first orientation angle dominates. The second harmonic is then at an orientation angle relative to the first harmonic. This is shown as a pseudo phasor diagram in figure 5.16. It should be noted that the rotating vectors cannot be added together to obtain an overall resultant vector as the second harmonic is rotating twice as fast as the first harmonic. It is also important to take into consideration that the rotating vector is not necessarily a straight line but of an elliptical shape.

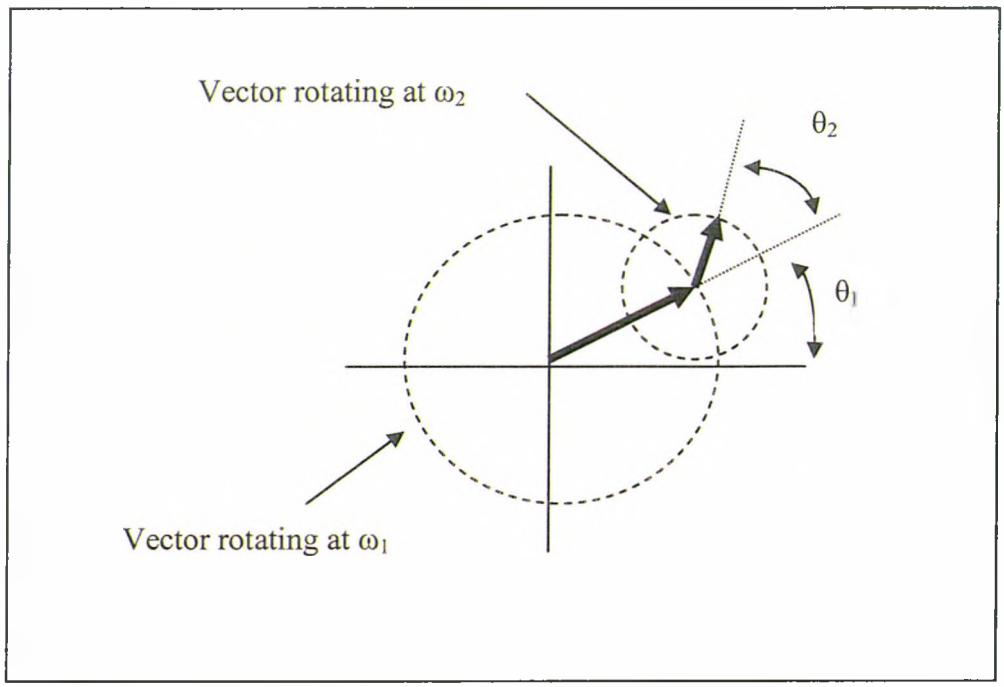


Figure 5.16 Diagram showing the 'addition' of two elliptical phasors rotating at ω_1 and ω_2 radians/sec and at orientations of θ_1 and θ_2

5.4. Interpretation of harmonic content with simulated gestures

Tests were carried out on a series of pseudo-trajectories. Some of the shapes were similar to those observed by Gibet (2001) and consist of straight line, curves (concave and convex), ellipses and other forms. These tests (Appendix V) enable the structure of the gesture to be interpreted from the harmonic content.

5.4.1. Planar triangular trajectory

A set of values were generated to form a 'triangular' trajectory in the time domain to mimic a simple gesture trajectory, starting from one coordinate increasing to another coordinate and then returning to the starting position, as shown in the 2D and 2DT view as shown in Figure 5.17. The coordinates were chosen to produce a straight line in 2D space that has an orientation to the 'x' axis. With 31 points defining the trajectory, Fourier analysis gave the results of Table 5.3. The results compare very favourably with that given by theory for the amplitude of the odd harmonics being proportional to $1/n^2$, for a triangular waveform. The theoretical amplitudes of the 3^{rd}, 5^{th} and 7^{th} harmonics are 0.111, 0.04 and 0.0204 respectively, showing less than 1% error in all cases. It is observed that there are no even harmonics generated, no phase shifts, and each harmonic starts in phase at the initial point. The only difference that these results show to the normal 1D signal analysis is that the plane or orientation of this waveform is at an angle of 23.43° where with real data it would be wholly in the x or y axis. The orientation angle is the same as that subtended in the spatial domain.

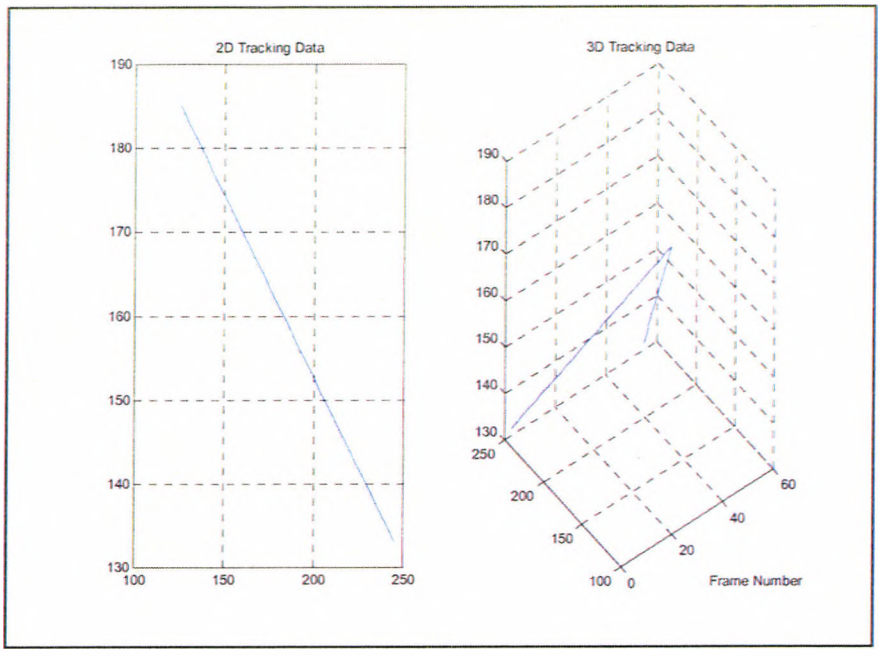


Figure 5.17 2D and 2DT profile of a 'Triangular' gesture trajectory

Harmonic	Positive Magnitude. A_p	Positive phase φ_p	Negative Magnitude A_n	Negative phase φ_n
1	1	23.4	1	23.4
2	0	23.4	0	23.4
3	0.11	23.4	0.11	23.4
4	0	23.4	0	23.4
5	0.04	23.4	0.04	23.4
6	0	23.4	0	23.4
7	0.02	23.4	0.02	23.4

Table 5.3 Triangular trajectory magnitude and phase information for the first 7 harmonics, based on 31 points

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	0	1	0	23.4
2	0	0	0	0	0
3	0.11	0	0.11	0	0
4	0	0	0	0	0
5	0.04	0	0.04	0	0
6	0	0	0	0	0
7	0.02	0	0.02	0	0

Table 5.4 Triangular trajectory magnitude and phase information based on 31 points

Refinements were made to the triangular waveform data to make it similar in amplitude to an actual gesture but following a 'triangular' time domain trajectory. This was undertaken to be the first step to simulate a gesture and obtain an appreciation of the type of harmonics that might be expected. The image of Figure 5.18 shows two images superimposed on each other, showing the hand at the start and at the height of a possible straight line trajectory. However, it is noted that the sudden change in direction of the trajectory, at the apex of the waveform, is unrealistic in the real situation but it does give an indication of the performance of the system.

In addition this trajectory shape was used to test the performance of the system to the time normalisation process to ascertain what effect it had on the harmonic analysis. Harmonic results were obtained which compared triangular trajectories based on a different number of original samples and then at different orientations. Changing the number samples tested the time normalisation process and the calculation of the L_1/M_1 and L_2/M_2 ratios and their interaction. Mean square errors of the harmonic magnitudes and phases did not show significant change in values between tests and so were a poor indicator of differences in performance. Close inspection of the overall system results showed some deviations in performance that was detectable mainly with phase differences. One of the worst cases of difference was with the waveform defined by 49 points. The results of harmonic analysis of this waveform are presented in Figure 5.19 and Table 5.5.

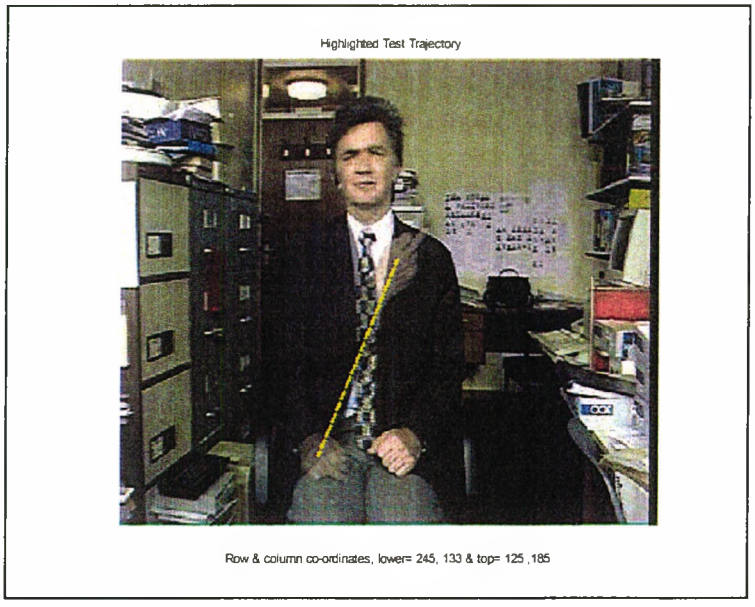


Figure 5.18 Formation of a triangular trajectory

All harmonics are in the same plane, as before, and with an orientation angle of 135 degree expected due to the x-y values. All positive and negative magnitudes are equal, so a straight-line trajectory results. The implementation of the L_1/M_1 and L_2/M_2 ratios gave a normalisation of 65 points instead of 64, even though the theoretical calculations showed negligible error. To compensate for this, the last point was truncated, and hence gave a small discontinuity in the periodic input waveform. The result of this is an expected phase shift of one 64^{th} in 360° i.e. 5.63° which is very similar to the result of 5.7° recorded. Some low level even harmonics, typically 4 to 5% on the 2^{nd} and 4^{th} harmonics, occur this time. Furthermore, there is

a 5% change on the 3rd harmonic and 2% change on the 5th harmonic. Inspection of the waveform resulting from the normalisation process showed that at some ratios a slight distortion to the triangular shape occurs, and as a result small amounts of distortion creep into the harmonic analysis. It is observed that the acute angle of the triangular waveform is much more exacting than would be expected in any gesturing situation where changes in direction generally take place less abruptly. Most gestures do not make an abrupt change of direction at the height of the gesture, but pause for a while before changing direction and so the slight degradation would not normally be noticeable. Experimentation into changing the orientation angle of the trajectory showed that the system could resolve about 4 degrees, or about one percent of angular movement. A worst case phase error would appear to be about 1.5%, or approximately 6° for the system.

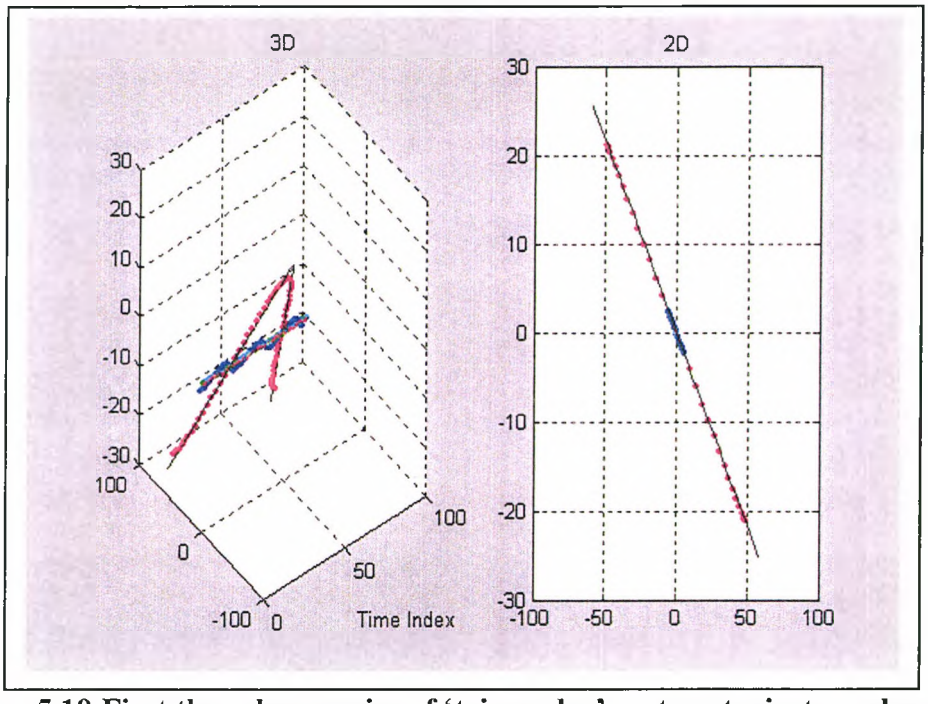


Figure 5.19 First three harmonics of 'triangular' gesture trajectory showing 2DT (left) and 2D (right) views all at an orientation equal to the spatial orientation angle (1st harmonic=red, 2nd harmonic=green, 3rd harmonic = blue)

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-5.7	1	5.7	135
2	0.043	-6.6	0.043	6.6	-180
3	0.16	-11.4	0.16	11.4	180
4	0.49	-26.5	0.49	26.5	-180
5	0.06	-34.2	0.06	34.2	180
6	0.012	9.3	0.012	-9.3	-180
7	0.034	-24.0	0.035	24.0	180

Table 5.5 Triangular trajectory magnitude and phase information based on original 49 points

5.4.2. Curved and Oscillatory Trajectories

A set of different possible trajectory paths was investigated, as detailed in Appendix V. These experiments were undertaken to gain insight into how the attributes of the harmonics changed with changes of shape in the trajectory. An example is shown in Figure 5.20 of an arc or shallow 'concave' trajectory, a trajectory that deviates from a straight line but where the rising and falling elements follow the same path in appearance space. The first four harmonics are shown as red, green, blue and cyan respectively. The corresponding harmonic values are shown in Table 5.6. The odd harmonic values are very similar to those of the 'triangular' gesture; the magnitudes decrease as the harmonics increase. However, the even harmonics now start to increase. The second harmonic is seen to be more prominent and at a different orientation angle to the plane of the odd harmonics.

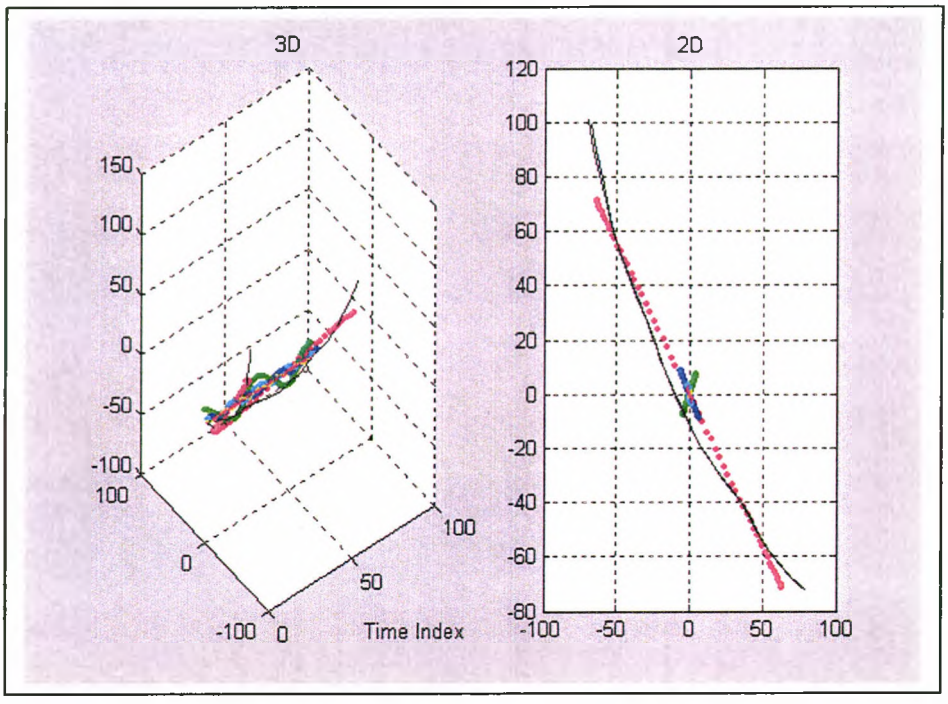


Figure 5.20 2DT and 2D views of the harmonics of a shallow concave trajectory showing 1st (red) and 3rd (blue) harmonics at the same orientation, but the 2nd (green) harmonic at a significantly different orientation

Another example of a deeper 'concave' trajectory shows the effect of the even harmonics to be more prominent. Changing the trajectory to a 'convex' shape gave similar magnitude values except that the phase of the orientation angle had changed by about 180°. It is interesting to note that the orientation angle for the 'triangular' and three curved trajectories, did not generate any ellipses as the 'rising' and 'falling' of the trajectories followed the same path. In addition, the orientation of the first harmonic remained relatively constant, ranging from -135° to -126° for data having similar sets of low and high trajectory coordinates.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-7.2	1.0	7.2	-131.8
2	0.093	-13.8	0.092	13.8	72.7
3	0.111	-20.7	0.113	20.7	-66.3
4	0.038	-24.7	0.038	24.7	23.4
5	0.037	-30.8	0.039	30.8	-14.2
6	0.023	-34.3	0.016	34.3	62.6
7	0.025	-41.5	0.031	41.5	-79.6

Table 5.6 Harmonic values of a shallow concave trajectory

Further experimental results are to be found in Appendix V. An elliptically shaped trajectory generated ellipses at most harmonics whilst staying mainly at the same orientation as the first harmonic. Deviations of harmonics from the first harmonic orientation were to compensate for the irregularities in the original trajectory shape. The 'figure of eight' trajectory has embedded into it two oscillations as can be seen by the second harmonic magnitude being significant with the orientation angle changing by about 90°. Another similar trajectory, but with six oscillations embedded in the trajectory, showed the expected prominence of the sixth harmonic.

5.5. Analysis and synthesis comparison

Validation of the positive and negative sequence modelling of the gesture trajectory can be accomplished by taking the harmonic components from the analysis and synthesising the waveform. It has already been noted that the first few low-order harmonics are generally sufficient to reconstruct the waveform to an acceptable level of similarity with the original waveform. For this exercise the 'figure of eight' trajectory has been chosen as the rising and falling trajectories cross emphasising the difference between this 2DT Fourier analysis and the Fourier Descriptor technique. In this case just the first, second and third harmonic components have been used. The first three harmonic coefficients, shown in table 5.7 are extracted from the full set shown in Appendix V. The coefficients that describe the harmonic being the positive and negative sequence amplitudes, the orientation angle θ and the phase angle ϕ.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	1.98	1.156	-1.98	-129.3
2	0.299	90	0.291	-90	89.3
3	0.146	4.8	0.073	-4.8	-75.2

Table 5.7 The first three harmonic coefficients used to synthesis the original waveform

The appearance-based picture (top-left) of Figure 5.21 shows the 'figure of eight' contour and also in the 2DT picture (bottom-left) although the starting and finishing locations can be more readily seen (starting location 'o' and finish location '*' in

blue). The second harmonic orientation angle at near 90° confirms the original experimentation that it is mainly responsible for the cross-over characteristic of the trajectory.

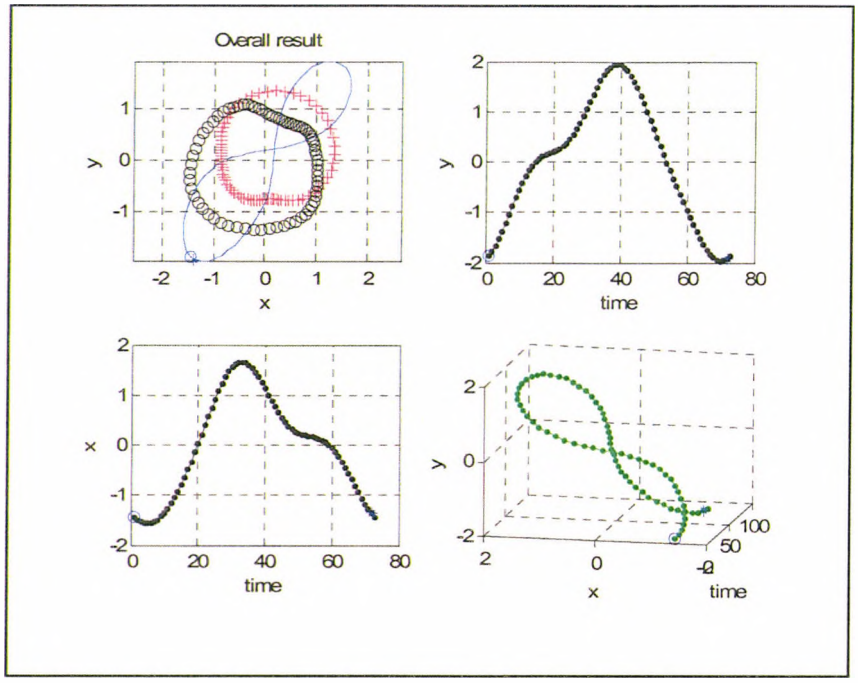


Figure 5.21 Four views of the synthesis of the 'figure of eight' trajectory from the coefficients of the first three harmonics given in Table 5.7

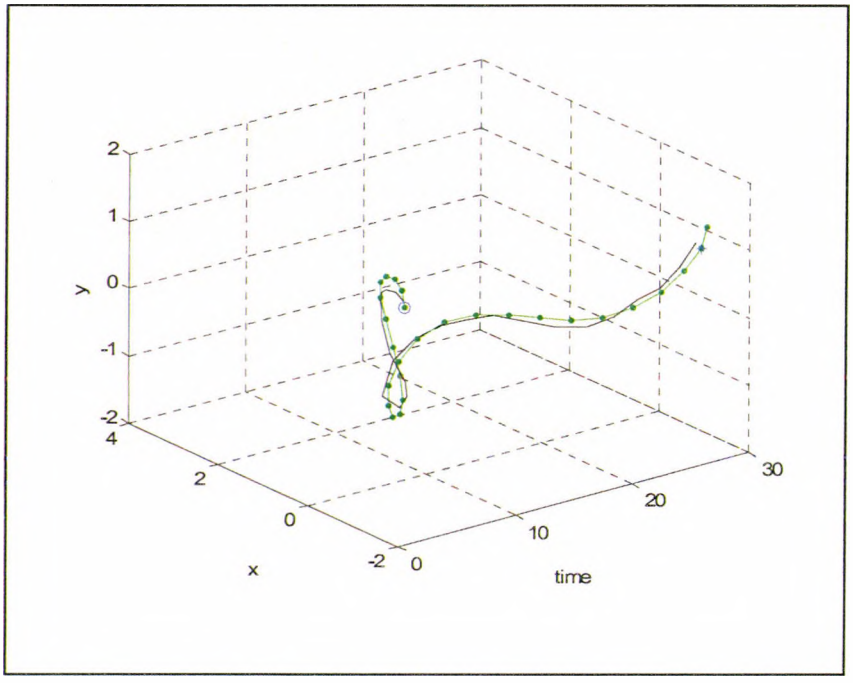


Figure 5.22 A 2DT comparison of original trajectory black, transposed and scaled, with the synthesised trajectory (green).

To be absolutely sure that the synthesised trajectory was similar to the original trajectory, scaling and transposition was undertaken. The offset of the original trajectory was adjusted so that it coincided with the offset (zero) of the simulated trajectory. The scale of the original trajectory was scaled so the maximum and minimum values were the same as the simulated trajectory. The results are shown in Figure 5.22 where the original trajectory (black) is compared with the synthesised trajectory (green) at a slightly different angle to the previous figure to show the trajectory changes over time more easily. It is apparent that the synthesised trajectory is a close match to the original trajectory albeit it is smoother, as just three harmonics are used.

5.6. Analysis of harmonic components from some gestures

A gesture trajectory showing arc type characteristics was evaluated (Appendix V). It is noticeable that the first harmonic generates an ellipse as the 'rising' and 'falling' paths are different. However, the second harmonic produces virtually a straight line in appearance-space with a significant deviation of the orientation angle from the first harmonic orientation angle. The third and fourth harmonics generate ellipses and the third harmonic ellipse is seen in Figure 5.23 as an 'elliptic corkscrew'.

An example of the harmonics generated for the data shown in Figure 5.28 of a hand raising gesture, from the PETS database, is given in Table 5.8. The gesture coordinates are shown as cyan coloured symbols 'o', in Figure 5.24. The harmonic content of this trajectory was found, and then just the first 6 harmonics used as input to the IFFT to reconstruct it. It is interesting to note how well the trajectory is represented by just six harmonics.

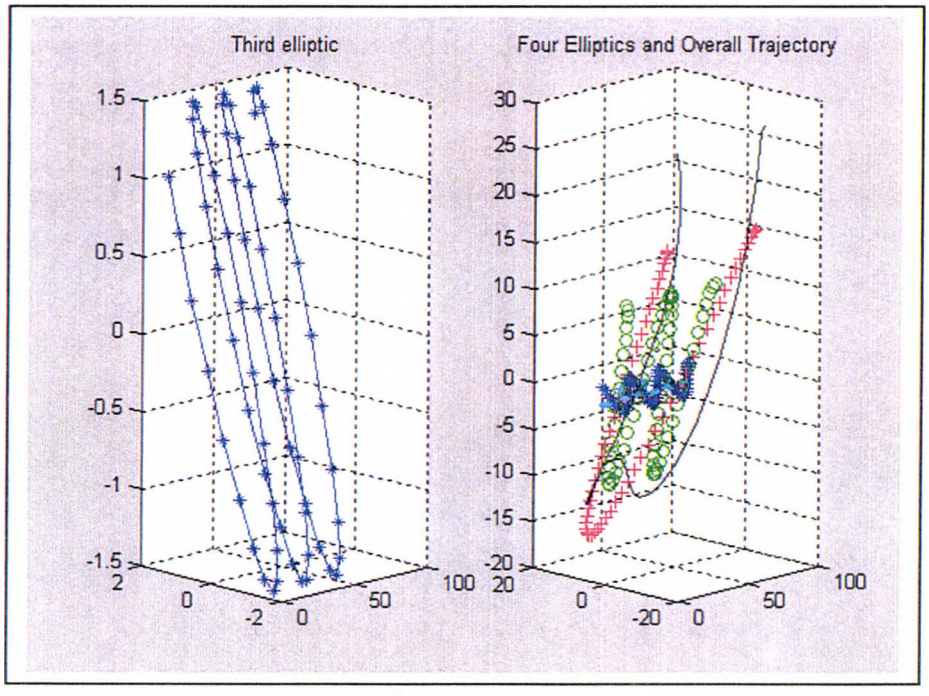


Figure 5.23 2DT views of an arc type gesture trajectory (black on right image) that has an arc or non-planar trajectory characteristic. The third harmonic demonstrates an 'elliptic corkscrew' (blue) as shown on the left-hand picture. The first (red) and second (green) harmonics are shown in the right image.

Harmonic	Positive Real Magnitude	Positive Imaginary Magnitude	Negative Real Magnitude	Negative Imaginary Magnitude
1	0.981	+0.195i	0.934	- 0.810i
2	0.402	- 0.236i	0.223	- 0.292i
3	0.208	- 0.206i	0.156	- 0.074i
4	0.042	- 0.069i	0.049	- 0.083i

Table 5.8 Table of complex data representing the first 4 harmonics of a gesture trajectory

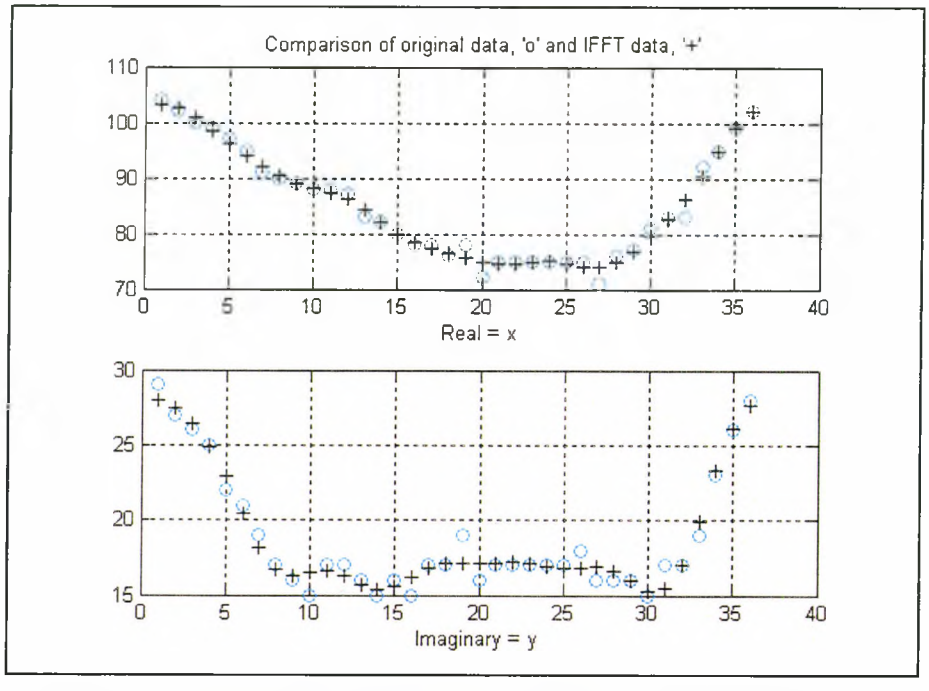


Figure 5.24 Original data (cyan circles) and data produced by the IFFT from the first 6 harmonics (black crosses)

The harmonic data produced by the FFT is in complex form, and a typical set of results for the first 4 harmonics are shown in Table 5.8. Very little meaningful data can be gained from this presentation apart from the magnitudes decreases as the harmonics increase. Further insight can be gained by converting the data into magnitude and phase format, as shown in Table 5.9. The results of Table 5.9 can alternatively be presented as a graph as shown in Figure 5.25. This graph compares the positive and negative magnitudes and phases about the central d.c. frequency. The positive sequence harmonics values increase from the centre to the right, whereas the negative sequence values increase from the centre to the left. This presentation of harmonic content allows a visual comparison and interpretation of magnitudes and phases of the positive and negative sequences.

The average phase of the positive and negative sequences gives the absolute orientation angle and is used in the next stage of presenting the data in a meaningful way. As discussed earlier, a more useful interpretation of the phase angle is by equating the positive and negative phase shifts to the orientation angle and the phase shift, where: -

$$\varphi_p = \theta + \phi \text{ and } \varphi_n = \theta - \phi$$

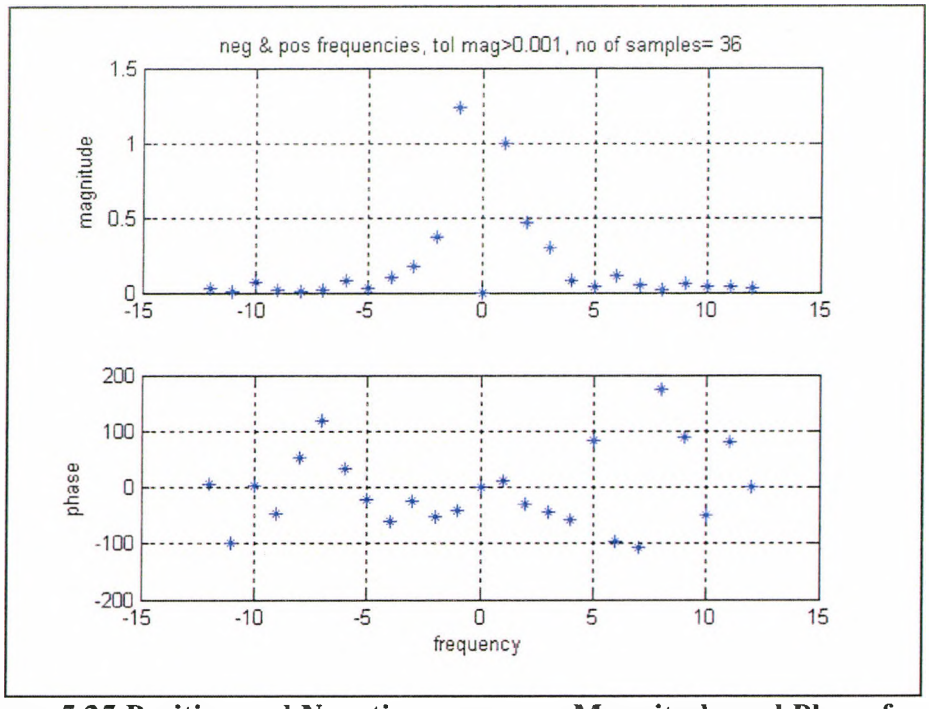


Figure 5.25 Positive and Negative sequences Magnitude and Phase for 12 harmonics

Harmonic	Positive Magnitude, A_p	Positive angle φ_p	Negative Magnitude, A_n	Negative angle φ_n	Average Phase (Absolute Orientation) θ
1	1	11.2	1.237	-40.9	-14.9
2	0.466	-30.4	0.368	-52.6	-41.5
3	0.293	-44.8	0.172	-25.4	-35.1
4	0.081	-58.6	0.096	-59.7	-59.2

Table 5.9 Table of magnitude and phase representing the first 4 harmonics of a gesture trajectory

An alternative presentation of the data is to arrange the phase in the form of phase angle and orientation angle, as shown in Table 5.10, where the phase shift of the positive and negative sequences is separated from the orientation angle. The orientation angle is affected by the previous orientation angle, as discussed in a previous section, so it is more convenient to express the orientation angles as to their relative orientation. In general it is observed that the orientation of the nth harmonic ellipse is: -

$$\theta_n = \theta_{A1} + \theta_{A2} \ldots . \theta_{An}$$

The phase and absolute orientation angles of Table 5.9 are adjusted to show phase shift and relative orientation angles of each harmonic in Table 5.10.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	26.1	1.237	-26.1	-14.9
2	0.466	11.1	0.368	-11.1	-26.6
3	0.293	-9.7	0.172	9.7	6.4
4	0.081	0.6	0.096	-0.6	-24.1

Table 5.10 Table of magnitude and orientation angle separated from phase for the first 4 harmonics of a gesture trajectory

Alternatively the phase information can be shown on an Argand diagram of Figure 5.26. The position of the rotating positive and negative sequence phasors are shown by the red '+' and black 'o'. When the phase shift ϕ is near zero (0.55°) as with the fourth harmonic the two symbols virtually overlap, and the resultant orientation angle, θ is the angle between this position and the origin. When the phase shift is greater than zero the positive and negative positions move further apart as can be clearly seen for the first harmonic.

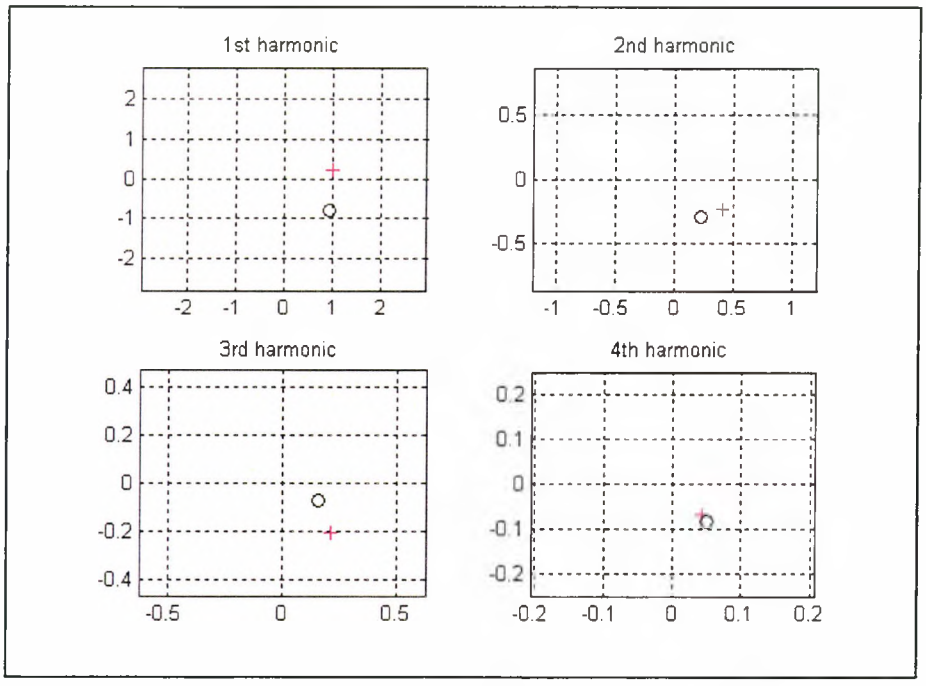


Figure 5.26 Positions of the first 4 harmonics on an Argand diagram for positive (red cross) and negative (black circle) sequence.

5.7. Performance Assessment

5.7.1. Gesture Truncation

It is noticeable that in real situations that the dominant gesturing hand does not necessarily return to the exact starting position. Investigations were carried on data of hand gesture waving from the PETS database. The approach taken is illustrated in Figure 5.27. The red dotted lines represent the search area for locating the first

motion of a gesture sequence. The cyan dotted line represents the updated start position and the magenta dotted line represents the stop condition.

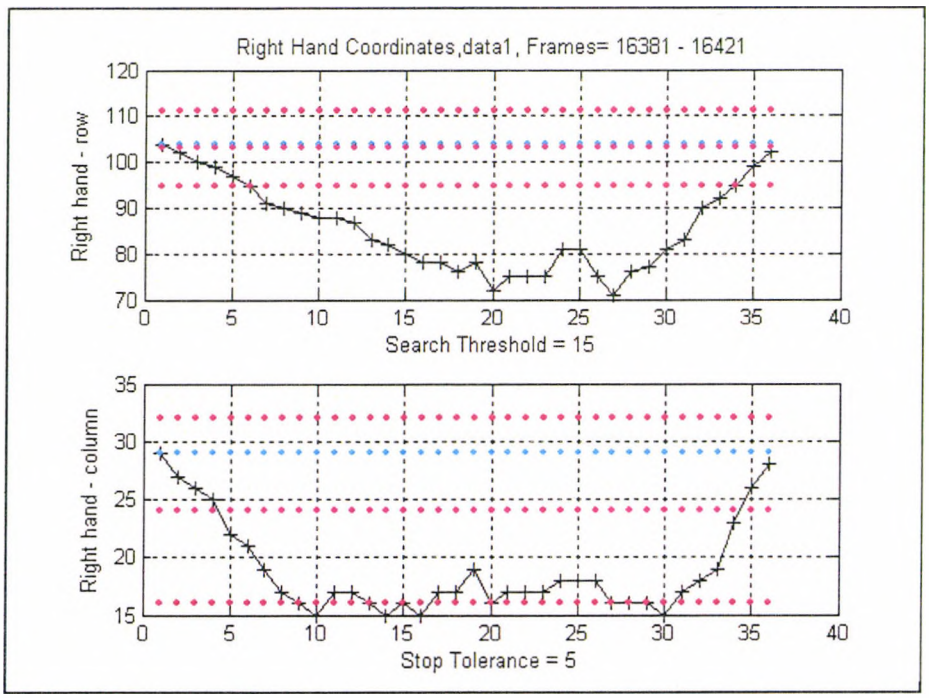


Figure 5.27 The Stopping Tolerance set to within 5 pixels of the Start Coordinates

The tests were carried out with stopping distances set at 2, 5, 10 and 15 on the row coordinates from the starting position. Inspection of the phase shift, ϕ of the first harmonic shows the phase shift changing by about 7° for every step that the sequence shortens. For example, when the gesture sequence reduces from 36 to 32 in length, the phase shift changes by 28° i.e. 7° per step. This change in phase is shown in Figure 5.28 and shows how the phase shift adjusts for the discontinuity in the waveform as the gesture becomes shorter. It can be seen that the coordinates for the four tests for each test form an arc centred at the origin, showing the phase shift that occurs.

However, more significantly, the orientation angle of the first harmonic remains virtually unchanged, varying only 4.4° between the four conditions. This is shown in the Figure 5.29 for the first 6 harmonics. It is noticeable how closely clustered all harmonics are for the four test conditions. Overall the phase shift approximates to $k\phi + \theta$, where 'k' is the harmonic, 'ϕ' the time domain phase-shift and 'θ' the orientation angle. There is some indication that at the greater stopping thresholds, some of the higher order harmonics start to increase due to 'spectral leakage' caused by the discontinuity. The seventh harmonic makes a steady increase from 0.042 to 0.15 as the stopping thresholds change from 2 to 15 pixels.

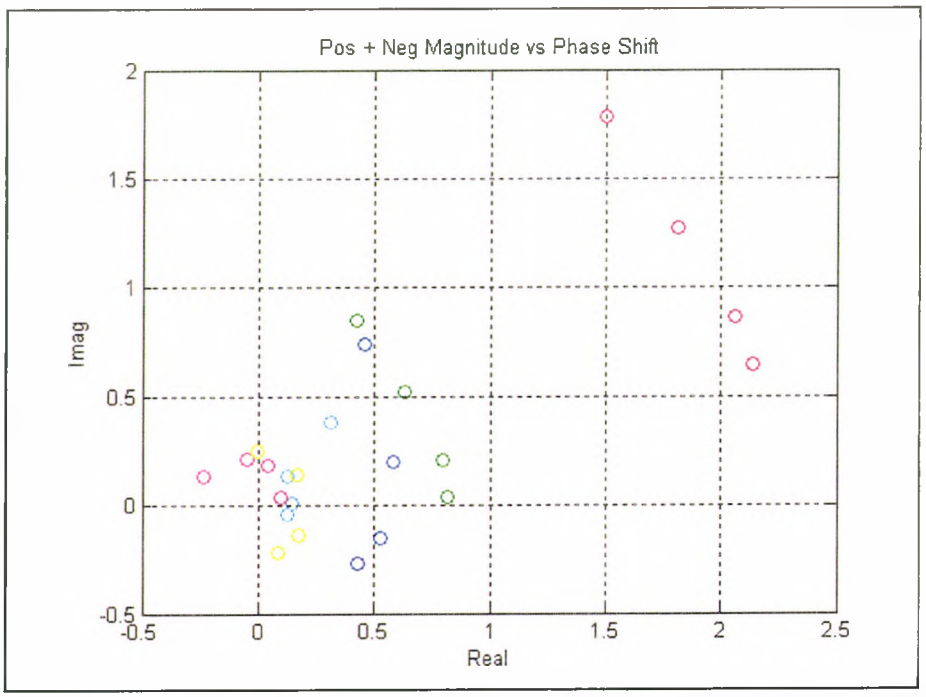


Figure 5.28 Argand Diagram representation of magnitude and phase shift for the first 6 harmonics, when the Stop Tolerance is 2, 5, 10, 15 (first to sixth harmonic, red, green, blue, cyan, magenta and yellow respectively).

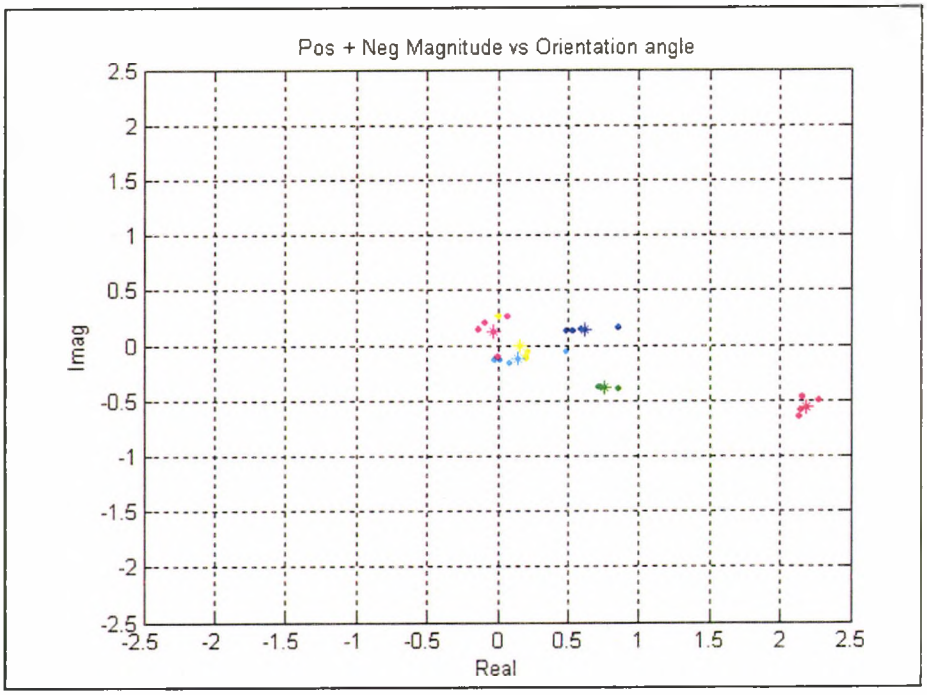


Figure 5.29 Argand Diagram representations of Magnitude and Orientation Angle (first to sixth harmonic, red, green, blue, cyan, magenta and yellow, respectively).

5.7.2. OSA Performance Based on Harmonic Analysis

The performance of the OSA can be assessed in a number of ways. The resulting trajectory can be compared, numerically or visually with the coordinates obtained manually/visually, as described in chapter 4. But the source of data to the OSA can vary because the input data can originate from SCM objects, SCMI objects, SCME objects or SCMEI objects. It has already been established that the latter objects do not perform very well in poor lighting conditions. However, in good lighting conditions all four methods perform well. There are less SCMI and SCMEI objects produced with the most significant object being appropriate for tracking the trajectory the majority of the time.

The OSA performance is dependent on the search threshold. If the search threshold is too small then the next frame's gesture objects will not be located. A valuable comparison can be made for an example based on SCM, SCMI and visual/manual tracked data. The SCM data was obtained, as shown in Figure 5.30 and compared with that visually recorded as shown in Figure 5.31. It can be seen that the overall shape is the same, but during the mid-phase, static part of the SCM trajectory is inherently noisy due to the object capturing process. However, the visually obtained data is smooth because of the inherent low pass filtering action in the data acquisition process.

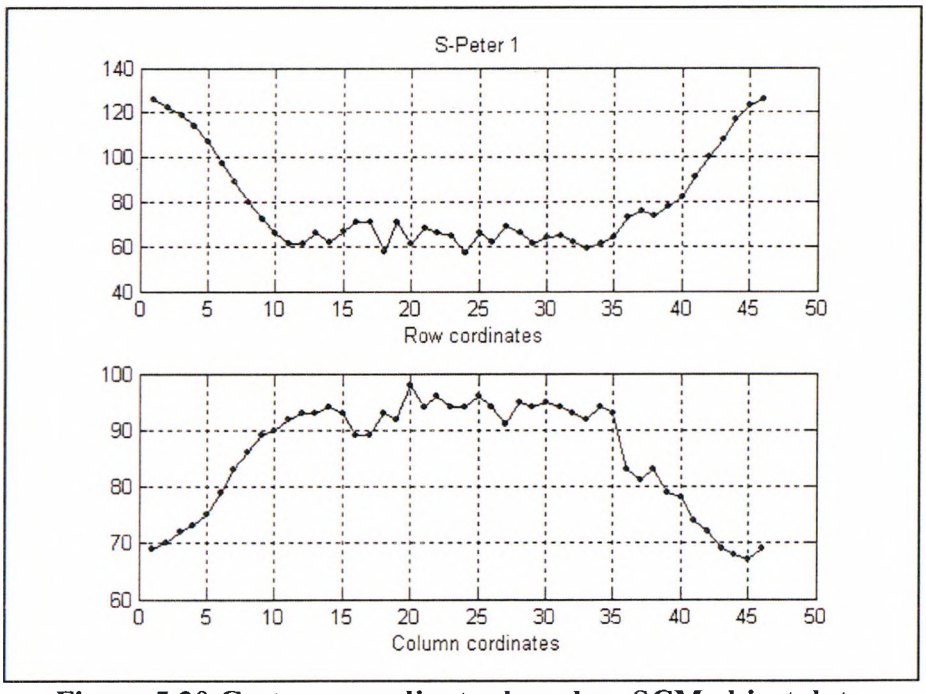


Figure 5.30 Gesture coordinates based on SCM object data

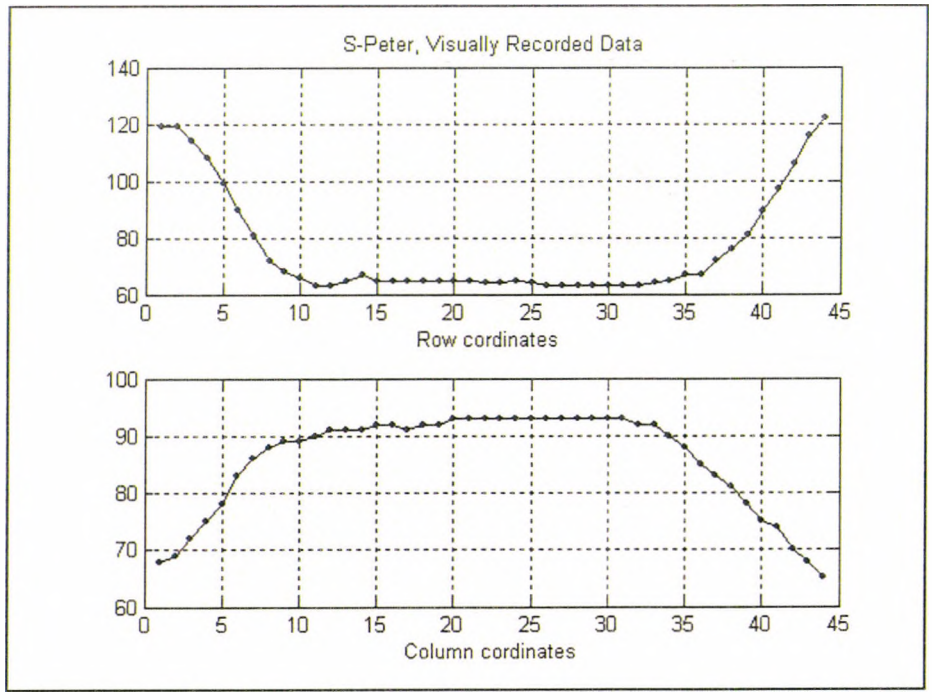


Figure 5.31 Gesture coordinates based on manually/visually recorded data

The OSA performance with the SCMI data between frames 8 to 12, as shown in Figure 5.32 did not follow the most significant object as it moved a distance outside the search zone. As a consequence the next four outputs use the previous output value and a plateau region can be observed in both the row and column coordinate values. The error in the OSA output occurred during the dynamic part of the trajectory, when the most significant object usually dominates other objects, in this particular case the velocity of the hand was larger than normal and exceeded the search threshold.

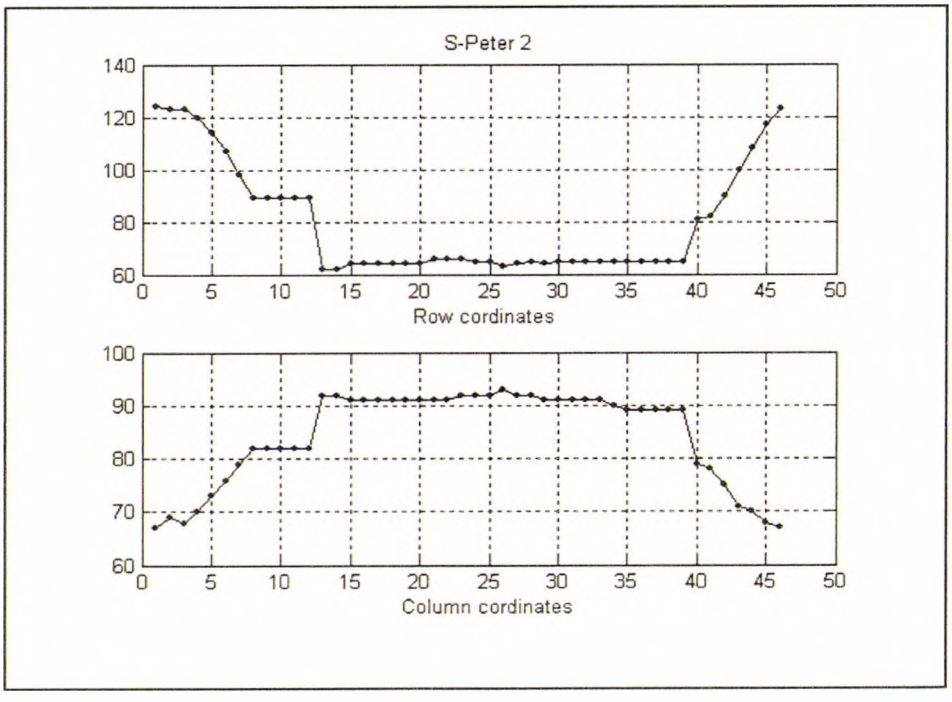


Figure 5.32 Gesture coordinates based on SCMI data

Application of the IFFT to the original input data (cyan coloured circles) to the OSA shows that using six harmonics is required to reconstruct the waveform to an acceptable shape (black crosses) as seen in Figure 5.33. The plateau region has now disappeared and for the remaining frames of the trajectory the input and reconstructed output data appear very similar. In addition the reconstructed data waveforms are compatible to that recorded using SCM objects or visually/manually recorded data. Using the low order harmonics introduces some smoothing to the waveform that allows degradation of the OSA's performance to be tolerated. It also removes the high frequency noise associated with error due to temporary tracking problems or the variability of the SCM objects position associated with the gesturing hand.

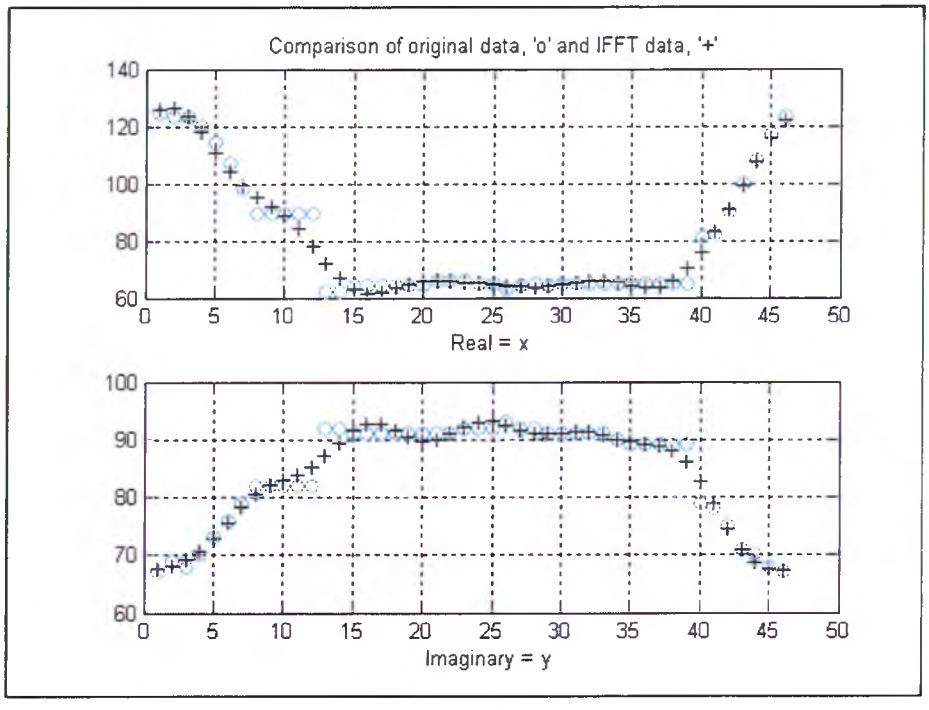


Figure 5.33 IFFT of SCMI trajectory data using 6 harmonics (black, '+') and original data (cyan, 'o')

All OSA tracking methods show close agreement of the magnitudes and orientation angles of the first 4 harmonics by used as detailed in Appendix V.

5.8. Pointing Gesture Experiments and Initial Interpretation

The pointing gesture experiments were devised to be similar to those conducted by Howell and Buxton (1998) in which four gestures were used; pointing the right hand to the left; pointing the right hand to the right; wave the right hand above the head (urgently) and wave the right hand below the head (not urgently). In this experiment (Harding and Ellis, 2004) used six subjects (A,B...F) who all undertook the same sequence of gestures. The subjects were seated on a chair with a web-cam used to record the five gestures that made up a gesture sequence, as detailed in Table 5.11.

Gesture	Action of right hand
1	To left shoulder and return
2	To the left and return
3	Straight up and down
4	To the right of subject and return
5	Straight ahead, 'halt' and return

Table 5.11 Gesture Number and Action

The trajectory that one person made doing five pointing gestures is shown in Figure 5.34. Each gesture was segmented and the harmonic data relating to that gesture produced using the Fourier analysis as previously described in this chapter.

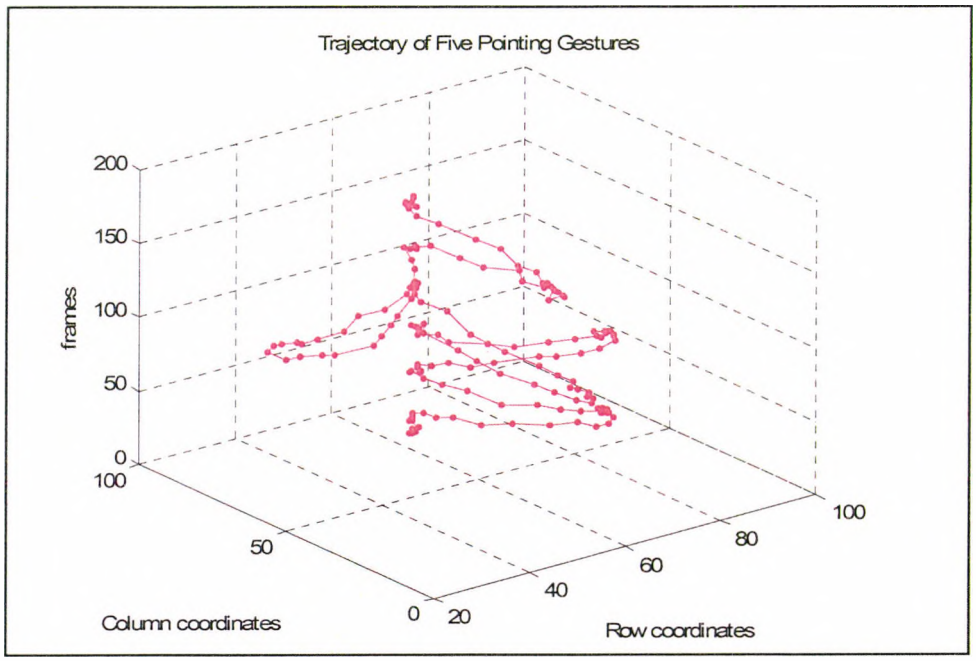


Figure 5.34 2DT view of the trajectory of five pointing gestures.

The four most significant harmonics of gesture A1 are recorded in Table 5.12. A comparison of the orientation angles produced by each gesturer for gesture 1, gave the values in table 5.13.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	3°	1.1	3°	-123°
2	0.317	6°	0.327	6°	26°
3	0.042	27°	0.103	27°	135°
4	0.027	68°	0.0513	68°	44°

Table 5.12 Orientation Angle, θ and Positive and Negative Sequence Magnitude and Phase, ϕ representation for gesture A1

Harmonic	Gesturer A	Gesturer B	Gesturer C	Gesturer D	Gesturer E	Gesturer F
1	-123°	-134°	-141°	-132°	-147°	-134°
2	26°	17°	25°	19°	23°	13°

Table 5.13 Comparison of all gesturers orientation angle for the first two harmonics of gesture 1

The first harmonic orientation angles for each gesturer show very similar values, averaging at –135° and deviates a maximum of plus/minus 12°. The second harmonics are also of similar values. The average value for the other gestures give first orientation angles that are distinctive of the gesture i.e. gesture 2, -170° and gesture 4, –21°. What this reveals is that the first orientation angle is related to the spatial coordinates of the gesturer. Although the physical dimensions of the gesturer may vary, and the interpretation of the gesture may vary, the orientation angle of each gesture is very closely clustered together.

Figure 5.35 shows the orientation angles and raw phase for four harmonics of the six gesturers for a particular gesture. In the majority of cases there is little difference between the raw phase shift calculated from the positive and negative sequence components and the orientation angle, because the phase shift, ϕ is very small as shown in table 5.12. This can be seen from the figure where the '.' represents the raw phase shift, and there is only one case where there is no overlap with the orientation angle. It is also evident that the second harmonic is even more tightly clustered, indicating the same curvature of trajectory from all gesturers. The average and standard deviation values for the first and second harmonics are 135° (std. = 8°) and 114° (std. = 9°), respectively.

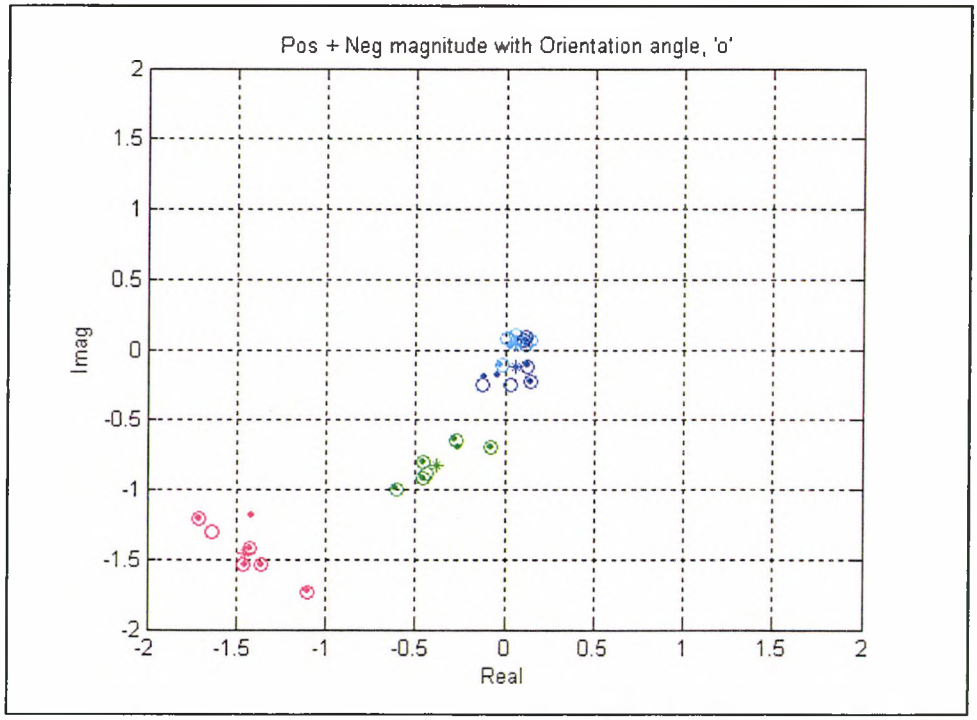


Figure 5.35 Orientation angle 'o', raw phase '.' , and average '*', for six gesturers performing the same gesture (red=1st, green=2nd, blue=3rd, cyan=fourth harmonic)

In the case of the first and second harmonics the data was closely clustered and average and standard deviation values were easily calculated. The values of orientation angles given for the third harmonic highlight some of the problems with the use of angles and the discontinuity at the 0°/360° or ±180° boundary. The orientation angles for the third harmonic are -45°, -83°, 32°, -59°, 38°and -116° giving an average of –39° and a standard deviation of 63°. The standard deviation measure cannot be relied on to give a useable answer when –45° could equally be represented as 315° and when not using negative numbers gives an average of 201° and standard deviation of 131°. Most realistically the problem with calculated values is likely to occur with clusters that straddle the ±180° boundary.

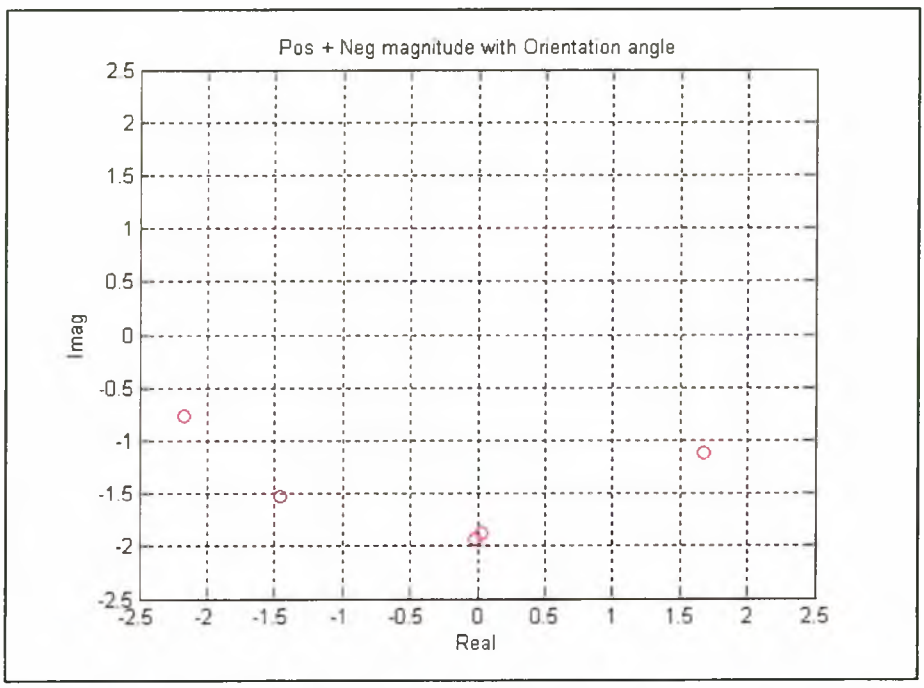


Figure 5.36 The first harmonic orientation angle for the five pointing gestures from one gesturer.

Most of the first orientation angles for the five pointing angles are clearly spaced, as shown in figure 5.36. The angles are –161°, -134°, –34°, –89° and –91°. The latter two are virtually the same as they represent the two gestures of 'Straight up and down' and 'Straight Ahead, halt and return'. In this case the contribution of the other harmonics is necessary to distinguish between the gestures. In this case the second harmonics are very similar at –78° and –95° but the third harmonics are quite different at 50° and –97°. A technique for distinguishing gestures by their harmonic content will be discussed in the next chapter.

5.9. Summary

The harmonic data, resulting from applying the FFT to the OSA data or manually obtained data, is in complex form. This form of data is not very meaningful, except the fact that as the harmonics increased the magnitudes generally decrease. Representing the data as positive and negative sequence components with

normalisation of magnitude and some manipulation of the phase data gives rise to phase information that can be used to characterise gesture trajectories.

Manipulation of the data into magnitude and phase information gave information about the structure of each harmonic, especially the phase. The phase information has two components. One component represented the phase shift of the harmonics in the time domain and the other phase component represented the orientation of the harmonic in the x-y plane or appearance domain. The orientation of any harmonic was found to be the sum of the orientation angles for all the preceding harmonics. In addition, the magnitude of the positive and negative sequence components indicated, if not equal, that an ellipse had be formed in the x-y plane and the direction of rotation depended on their relative magnitude. Rotation of the ellipse was anti-clockwise when the positive sequence magnitude was greatest and clockwise when the negative sequence magnitude was greatest. Viewing the harmonics in 2DT showed the ellipse could be pictured in time as an 'elliptical corkscrew'.

The first harmonic would revolve just once whereas the second harmonic revolves twice etc. The nature of each harmonic component gives valuable insight into the characteristic of the gesture. The first harmonic orientation angle gives an overall indication of the gesture direction. If the remaining harmonics have zero orientation angle then the gesture trajectory represents a straight line, in that the rising and falling parts of the trajectory take the same path. The second harmonic, and to lesser extent the other even order harmonics, is mainly responsible for determining the amount of curvature in the trajectory because the orientation angle is in a different plane to the first harmonic. When the magnitude of the positive and negative sequence components is different, an ellipse structure is indicated as the rising and falling paths of the trajectory are different.

Generally, the higher the order of the magnitude of the harmonic the smaller is the amplitude. This rule is broken if significant oscillation is detected or the even order harmonics are indicating significant curvature in the trajectory.

A number of simulated trajectories showed some other characteristics. If the rising and falling parts of the trajectory were the same, whether straight line or curved, then no ellipses were formed. However, with curved trajectories the even order harmonics, especially the dominant second order harmonic, moved out of the orientation plane of the first harmonics. If the rising and falling trajectories took a different path, then ellipses are formed as can be detected by the difference in magnitude between the positive and negative sequence magnitudes. Oscillation in a trajectory could also be detected. It could be envisaged that a second harmonic could result in either a curved trajectory or an oscillation in the trajectory.

Various simulated trajectories were devised to test the performance and characteristics of the system. A perfect triangular waveform showed that there were no time domain phase shifts and that the orientation angle was as expected by the data in the x-y plane. The magnitudes of each harmonic were as expected by theory. The length of the 'triangular' gesture was varied so as to ascertain the performance of the time-normalisation system (chapter 4). In some instances the interpolation and decimation process could not deliver the target length of 64 samples. Hence, the last value of the 65 length data was truncated. This truncation caused a small discontinuity in the waveform and resulted in a phase shift of one sixty-fourth of

360° affecting the first harmonic and higher harmonics at integer values of this phase shift. In addition, a small amount of distortion in the data was seen when the data was reconstructed with the IFFT. However, this distortion was considered to have minimal effect on real gesture trajectory data, as a hand movement was very unlikely to make the fast transition in direction as modelled by the triangular wave shape.

In this application there is a slight uncertainty as to where a gesture begins and ends. The beginning of a trajectory is dependent on detecting movement, which can be variable in speed, and hence the time elapsed for detection, albeit relatively small. The variability of speed will result in small changes of phase from trajectory to trajectory. In addition, there is the case of finding when the gesture stops. Observations had shown that the gesturing hand does not always return to the starting position. Investigations of allowing the stop criteria to range a number of distances from the starting coordinates gave rise to truncation errors, which showed up as time domain phase-shifts. In general these phase shifts due to truncation error only produced an error of about 1-2% error. But even with these changes of phase-shift the orientation angle remains tightly clustered for all low order harmonics and appear to be an invariant property of this analysis.

Assessing the performance of the OSA to different data sources showed that sometimes its performance was not ideal. Data produced visually/manually and from SCM and SCMI objects showed subtle differences. In one particular case, using SCMI data there appeared to be a distortion in the tracking process. However, reconstruction of the waveform with just six harmonics showed that the anomaly had been smoothed out. The frequency analysis showed that the orientation angles stayed virtually invariant to changes or errors in the capture or tracking process prior to transformation into the frequency domain and so important information about the characteristic of the gesture is not lost.

It was found that the positive and negative sequence equations gave the same result as the more traditional matrix form of equations for describing elliptical structures. This was particularly useful as the equations were then in the same form as the results given by Fourier analysis. Syntheses of gesture trajectories were able to be obtained using just three harmonics. The analysis has shown that gesture trajectories can be characterised by the magnitudes of the positive and negative sequence components; the phase shift and the orientation angle of each harmonic. The next two chapters are concerned with the recognition performance of Probabilistic Neural Networks (PNN) using the data and properties of the harmonic components in association with clustering techniques to select target gestures.

6. Gesture Recognition using Probabilistic Neural Networks and Hierarchical Cluster Techniques

This chapter investigates techniques to classify gestures. In the previous chapter it was established that the orientation angle of the first harmonic played a major role in classifying the pointing gesture. It has also been shown that just a few low order harmonics adequately characterise a gesture. The classification technique uses a vector based on the magnitude of the sum of the positive and negative sequence components (the length of the major axis of the ellipse) and the orientation angle of a few harmonics. The coordinates of the vector are used in cluster analysis to find a target gesture for a particular gesture. The target gestures are then used with a Probabilistic Neural Network (PNN) to identify or analyse unknown gestures

6.1. Introduction

The technique introduced in this chapter is similar but different to the object recognition technique based on Fourier Descriptors (Kuhl et al. 1982, Lin et al. 1987 & 1990,) using spatial sampling. With this method the magnitudes of harmonic components are used to identify shape. Comparing the harmonics of an object with a number of reference shapes in a database allows object recognition to be achieved. A two-dimensional array of coefficients, one row for each object, is stored in the memory and used for later comparison. The nearest neighbour distance between the object and each reference in the database is used to decide whether the unknown shape is similar to the database shape or not. Typically a calculation, as shown below, is implemented: -

$$D = \sum_{i=1}^{k} (a(i)_{ref} - a(i)_{OBJECT})^2$$

The square difference between the magnitudes of the harmonics, a(i), are calculated and the sum 'D' of all the harmonics, 'k' are found. If the value 'D' is smaller than a predefined threshold for an entry in the database, then the actual shape is regarded as recognised.

Establishing a target gesture in 2DT space is more complex than the case for identifying non-deformable objects. But the least squares formula only considers the magnitude of the harmonics and does not include phase information which is an important parameter in characterising a gesture. The phase information, in the form of complex data representation can be used with a PNN.

In addition, the PNN is an appropriate tool for gesture recognition as it is structured to take the harmonic, multidimensional inputs. It is also superior to least squares methods where outliers can cause errors that produce false results (Bishop, 1995). For example, an unknown gesture is input to a PNN and compared with a number of Target gestures to establish which one it is closest to. The Target gestures are established by clustering techniques. Because the amount of data for these experiments is generally sparse, hierarchical clustering techniques were investigated rather than the more common k-means. PNN are generally constructed from Radial Basis Functions. They are seen as ideal for practical pixel-based vision applications

(Howell & Buxton, 1995) as they are particularly efficient at processing sparse, high-dimensional data (which is common in images) and because they use approximation, which is better than interpolation for handling 'noisy', real-life data. They also provide a guaranteed, globally optimal solution via simple linear optimisation.

In this chapter the PNN networks are applied to the classification of five pointing type gestures. The performance is evaluated for different target gestures and the number of harmonics used. The influence of the first harmonic is investigated and applied in another commonly used classification technique using a clustering algorithm. Hierarchical clustering techniques are also investigated using tools in the Matlab Statistics Toolbox. In this investigation different distance metrics and linkage methods are assessed. Furthermore, a comparison is made of the PNN and clustering techniques and ideas are presented for using the two techniques together for more complicated gestures to analyse in the following chapter.

6.2. Probabilistic Neural Network

There are many papers and reports covering all aspects of Artificial Neural Networks (ANNs) and gesture recognition systems, but very few on the employment of the Radial Basis Function to this type of application. But there is a very well documented area of research being undertaken for the RBF used for phoneme recognition (Berthold, 1994) and for face classification techniques (Howell & Buxton, 1995).

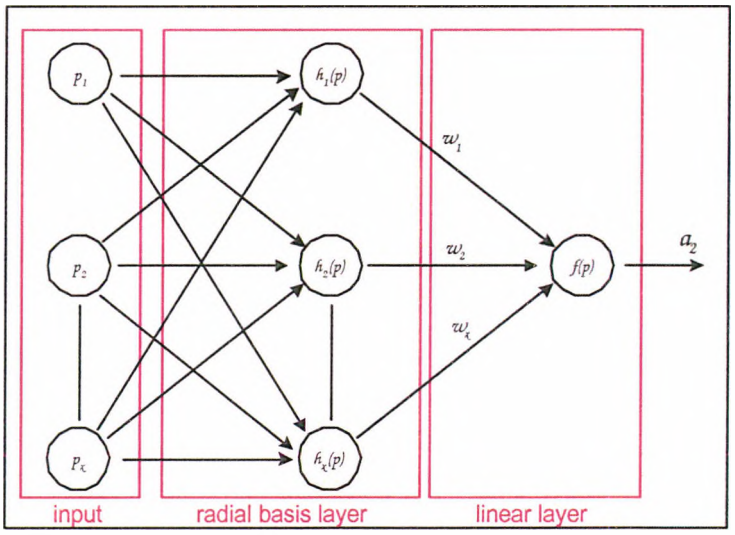


Figure 6.1 Schematic diagram of the RBF neural network, with the input layer to the left, RBF hidden layer in the middle and linearly combined with weights to give output f(x) to the right. (Source: Orr, 1996)

The Radial Basis Function was first used in neural networks by Broomhead & Lowe (1988). Other contributions to the theory include Moody & Darken (1989), Renals (1989) and Poggio & Girso (1990). The radial basis function (RBF) is a multi-layer, feed-forward network consisting of three layers.

Layer one is the input layer, layer two a hidden layer, which contains the basis function and layer three, the output which is linear. The architecture for the RBF network can be seen in figure 6.1. The neurons in the RBF have localised receptive fields because they only respond to inputs, which are close to their centres. This is in contrast to normal multi-layer perceptrons which have a global response to inputs due to the normal use of a sigmoid transfer function. (Hagan et al., 1996).

Layer one contains the inputs to the network, p_1 to p_x. In layer two the neurons in the second layer contain the RBF; a statistical transformation is based on (but not limited to), the Gaussian distribution.

$$z_j(x) = \exp(-\frac{|x-\mu_j|^2}{2\sigma_j^2}$$

where, σ^2 and μ_j are the standard deviation and the mean of the j^{th} unit receptive field. The output from layer one is the distance between the input to the network and the centre of the basis function, with the position and width of the function being learnt from training data. This is the characteristic feature of the radial function - response to an input will decrease or increase monotonically with distance from a central point, i.e. as the input moves away from this centre point, the neuron output will rapidly drop towards zero. A different centre point is calculated for each input into the network. The radial basis function is of the form that the output is a maximum of one when the difference between the input and the weights is zero. The output decreases to 0.5 when the difference measure increases to ±0.833.

The Matlab Neural Network Toolbox shows that the sensitivity of the radial basis neuron can be adjusted by the bias, b that is part of the layer two. If the bias is changed from 1 to 0.1 the spread of the input vector changes from 0.833 to 8.33 for an output of 0.5. In Matlab a critical mathematical scalar value is that of 'spread'. This value is set at design time and is concerned with the second layer, the RBF layer of the network. If a neuron's weight vector is a distance of 'spread' from the input vector, its weighted input will be spread, its net input will be (-log(0.5) = 0.833), making its output 0.5. Each bias in layer two is set to 0.8326/spread, which gives a radial basis function which cross at 0.5 at weighted inputs of ±spread. This is the width of an area in the input space to which each neuron will respond. For example, if 'spread' is set to 4, then each neuron in the RBF layer will respond with 0.5 or more to any input vectors within a vector distance of 4 from the weight vector. The value of 'spread' should ideally be large enough that neurons respond strongly to overlapping regions of the input space.

There are several sub-classes of the RBF network, one of which is of particular value to this project. It is the Probabilistic Neural Network (PNN). The PNN is a three-layer network used for pattern classification type problems. Layer one, as always, is the input layer. Layer two contains the RBF layer whilst layer three contains a competitive layer. This type of network requires the prototype input pattern to be known and incorporated into the network as rows of a weight matrix. The competitive layer is so named because each single neuron in the layer 'excites' itself and inhibits all other neurons. The second layer sums the contributions from the first layer. Each class of input produces as its net output a vector of probabilities. The

final layer picks the maximum of these probabilities to produces one for that class and a zero for other classes.

The approach taken in this thesis is to use frequency domain data as input to the neural network rather than time domain data. Masters (1994) found that complex-domain neural networks are especially superior to real-domain networks when the information relevant to the solution is primarily embodied in phase relationships. He found that there were two neural network models to be particularly effective for supervised training, the Multi-Layer Feed-forward Network (MLFN) and the Probabilistic Neural Network (PNN). Masters used frequency data as input to these neural networks reporting that the main disadvantage to the MFLN is that excruciating long training periods are often required. Furthermore, Masters states that the execution time of a PNN is very slow and requires a relatively large amount of memory. However processing speed of computers has increased significantly since Masters made this comment. The PNN has been found to have faster execution times, and as such is an appropriate classifier for gesture recognition application.

The advantage of the PNN network using the Gaussian RBF is that scaling is not required. The length of the input vector to the PNN is proportional to the number of dimensions and is suitably scaled. In this gesture recognition technique the number of dimensions is equal to the number of harmonics. Typically, the higher the harmonic the lower will be its amplitude. By definition the unknown vector will be of similar amplitude to the target value. The neuron for a particular harmonic will be a calculation of the distance of the unknown input from the target value. PNN networks also avoid the problem of least squares systems (Bishop, 1995) in which one of the difficulties of the standard sum-of–squares error is that it receives the largest contribution from points that have the largest error. Outliers with PNN networks have errors that tend to zero and so do not dominate the error term. The role of outliers is again a concern with clustering techniques when they are used to identify target gestures for use with the PNN

6.3. Clustering

Clustering is used in many fields relating to pattern recognition. Clustering uses a range of techniques to classify similar objects into different groups or to partition objects into clusters or subsets so that they ideally share some common characteristics. Jain et al. (1999) have reviewed pattern clustering methods from a statistical pattern recognition perspective and present taxonomy of clustering techniques showing that clustering techniques are principally divided into either hierarchical or partitional clustering methods.

6.3.1. Distance Metrics

Cluster analysis is a way to segment a set of objects into clusters of objects that are very similar although the profiles of objects in different clusters are quite different. The basic procedure of cluster analysis is to find the distance between every pair of objects in the dataset. There are a number of different distance measures that can be used. The City Block technique avoids the mathematical intensity of the popular Euclidean measure and the Minkowski metric generalises these first two metrics

(Moon and Stirling, 2000), whereas the Mahalanobis metric is a measure that may be more appropriate where probability contours are elliptical. The axes of the ellipse are parallel to the x and y axes. This model assumes that the feature values are still independent but the standard deviations; 'σ' is now different. The distance metric is defined as: -

$$d^2 = \frac{x^2}{\sigma_x^2} + \frac{y^2}{\sigma_y^2}$$

Euclidean distance metrics works well when the data set has 'compact' or 'isolated' clusters (Mao and Jain, 1996) but the drawback to the use of Minkowski metrics is the tendency of large-scaled features to dominate others. Bishop (1995) explains that one of the difficulties of the standard sum-of–squares error is that it receives the largest contribution from points that have the largest error. Long tails to a distribution can cause a solution to be dominated by a very small number of 'outliers', which have a large error. Similarly, incorrectly labelled data can produce a large error. The Minkowski-R error with values of R less than 2 reduces the sensitivity to 'outliers'. When R=1, the minimum error solution computes the conditional median of the data rather than the conditional mean. Yam et al (2002) compare the recognition rates for walking and running after clustering of the PWM (Phase-Weight Magnitude) feature and compares Euclidean and Mahalanobis distances.

Duda (2006) explains that the use of the Mahalanobis metric removes several of the limitations of the Euclidean metric:

1. It automatically accounts for the scaling of the coordinates
2. It corrects for correlation between the different features
3. It can provide curved as well as linear decision boundaries

However, there is a consequence to these advantages. The covariance matrix can be difficult to determine accurately. The memory and time requirements grow quadratically rather than linearly with the number of features. These problems may be insignificant when only minimal features are needed, but can become significant when the number of features increase.

Duda further suggests that if a simple minimum-distance classifier is satisfactory, there is no reason to use anything more complicated. However, if the classifier makes too many errors the possible reasons are that:-

1. The features may be inadequate to distinguish between different classes
2. The features may be highly correlated
3. The decision boundary may have to be curved
4. There may be distinct sub-classes in the data
5. The feature space may be simply too complex

6.3.2. Partitioning methods

There are many different partitioning methods (Jain, 1999) including graph theoretic, mixture resolving, mode seeking and square error of which the latter is a form of the popular k-means technique attributable to McQueen (1967). Moon and Stirling (2000), state that the k-means technique has been extensively used for clustering applications. A relevant reference within the context of signal processing is Linde et al. (1980). k-means clustering partitions objects into k mutually exclusive clusters so that each object within a cluster is as close to each other as possible, and as far away as possible from other clusters. Each cluster is characterised by its centroid or centre point. The number of clusters that is expected is set in advance although the training is unsupervised. This technique has been found to be appropriate for large data sets. Its main advantage has been its simplicity and speed. The disadvantage of this technique is that the results may not be consistent because of the initial random assignments of centres for several runs of the algorithm.

Hierarchical clustering investigates groupings in data over a variety of scales of distance, by creating a cluster tree. The tree is not a single set of clusters but rather a multi-level hierarchy. Clusters at one level are joined as clusters at the next higher level. This allows a decision to be made about the scale or level of clustering that is most appropriate for the application. Hierarchical clustering use a variety of linking techniques. Many of the linking techniques are variants of the single-link (nearest neighbour) technique (Sneath and Sokal, 1973). Other techniques are commonly used, for example, complete-link (largest distance) (King, 1967) and minimum-variance (Ward, 1963; Mutagh, 1984). The mechanism of the linking process is explained in Appendix VI. Jain et al. discuss some of the characteristics of the different linking schemes. The complete-link algorithm produces tightly bound or compact clusters (Baeza-Yates, 1992). The single-link algorithm suffers from a chaining effect (Nagy, 1968) and has a tendency to produce clusters that are straggly or elongated. However, the single-link algorithm can be more versatile and was shown to extract concentric clusters whereas the complete-link cannot.

6.3.3. Cluster validation

Jain et al. (1999) make the important point that all clustering algorithms will, when presented with data, produce clusters regardless of whether the data contains clusters or not. A further point is made about cluster validity and assessing the clustering procedure's output. The analysis often uses a specific criterion of optimality although the criteria are arrived at subjectively and there is little in the way of 'gold standards' in clustering except in well-prescribed domains. There are three types of validation studies possible: an *external* assessment of validity compares the recovered structure to an *a priori* structure; an *internal* examination of validity tries to determine if the structure is intrinsically appropriate for the data; and a *relative* test that measures two structures and measures their relative merit.

However an extensive range of options for clustering is to be found in the Matlab Statistics Toolbox (Appendix VI). This toolbox has extensive range of distance measures and linkage methods. When the distance measures have been calculated, the objects are linked together on a basis of closest proximity to form a binary,

hierarchical cluster tree. As the objects are paired together, the newly formed clusters are grouped together into larger clusters until a hierarchical tree is formed. This information can be conveniently illustrated by a dendrogram that plots the hierarchical information as a graph. The hierarchical information can be used as a visual aid to create clusters by either detecting natural groupings or by cutting off the hierarchical tree at arbitrary points.

Furthermore, to validate the cluster information a correlation between the original distance measures and the linkage can be made. If the clustering is valid there should be a high correlation, which can also be used to assess the different distance measures used. Linkage can be undertaken by a range of methods, i.e. 'single', 'complete', 'average', centroid' and 'ward' (Appendix VI). In order to determine natural cluster divisions in a dataset, the length of each link in a cluster tree can be compared with neighbouring links below it in the tree. If the link is approximately the same length as a neighbouring link there are similarities between the objects joined at this level and are said to exhibit a high degree of consistency. When there is a difference between the links at the same level there are dissimilarities and the link is said to be inconsistent with the links around it.

The inconsistent links can indicate the border of a natural division in a dataset. The 'inconsistency coefficient' is a measure of quantifying the relative consistency of each link in the hierarchical cluster tree. The coefficient is produced by comparing the length of a link in a cluster hierarchy with the average length of neighbouring links. A low consistency coefficient indicates that the object is consistent with neighbouring objects whereas a high coefficient indicates the object is inconsistent with those around it. The use of the coefficient and visual methods helps to resolve which distance metric and which linkage method is appropriate for the application.

Most explanations of clustering are based on two-dimensional feature vectors that represent a pair of coordinates and hence are easy to demonstrate graphically. However, clustering can be achieved for higher-dimensional data, but the issue of scaling or normalisation must be addressed. If scaling is not addressed one of the dimensions may have an undue influence on the clustering results. Cluster analysis can be a tool to help find similarities, but appropriate scaling is needed to condition the data for valid results.

6.4. Recognition of Pointing Gestures

The following experimentation investigated how the PNN and the clustering techniques were able to recognise gestures based on harmonic data.

6.4.1. Normalisation of frequency data

First and foremost both PNN and clustering data should be normalised to avoid scaling variations. Data for the PNN application normalised data on the basis of the value of A_p or the greater value of A_p or A_n of the first harmonic. Clustering normalisation is similarly undertaken but based on the coordinates of a vector representing the orientation angles of the harmonics. In chapter 5, it was explained that a signal of unit amplitude and angular frequency ω, can also be represented by

the double-sided spectrum of a positive and negative frequency component of amplitude of half the single sided representation. For ease of experimental analysis it is convenient to designate both the positive and negative sequence components to unity, for real data, and hence making the single-sided magnitude equal to 2.

When the magnitude of the positive and negative sequence components have a magnitude of unity, a straight-line of value 2 is obtained in appearance-based space. This is a special case of an ellipse with major-axis, A equal to 2 and minor-axis, B equal to 0. This is because $Ap_k + An_k = A$ and $Ap_k - An_k = B$. For other cases where Ap_k and An_k are not equal an ellipse is formed but the major axis is less than 2.

In the pointing experiments, data obtained from the first harmonic, is often in the form indicating a shallow ellipse, and so positive and negative sequence components have very similar magnitudes of around unity. As a first approximation to normalising all gestures for scale it was decided to arbitrarily normalise all harmonic components by Ap_k for the first harmonic as the major axis of the ellipse will be close to 2 in most cases. However, in cases when the first harmonic is noticeably elliptical errors will creep into the analysis of first and subsequent harmonics. To remedy this situation it is convenient to normalise harmonics to the greater magnitude of Ap_k or An_k. This then also conveniently indicates if the direction of revolution of the ellipse is anti-clockwise or clockwise, respectively. It is appropriate to normalise all gestures to the length of the major axis of the first harmonic which has a maximum value of two. So for example when Ap_k is 1 and An_k is 0.8 , the major-axis length is 1.8, so the normalising ratio for all harmonics from a particular gesture are multiplied by the ratio of 2/1.8.

A vector is generated from the normalised major-axis magnitude and the orientation angle for each harmonic, and corresponds to visual observation of 2D views of harmonics shown in chapter 5. The coordinates of the vector are then used for cluster analysis.

6.4.2. Recognition using the PNN

The harmonic data, from the experiments (Appendix VII) with the pointing gestures as discussed in Chapter 5 were arranged in a form suitable for the PNN. The data was used in its complex data form, rather than magnitude and phase form to avoid discontinuity problems with the phase information (Chapter 5). The harmonic data consist of contributions from the positive and negative sequences, and each has real and imaginary values. Data from each harmonic is offered to the PNN as 4-tuples (the real and imaginary number of the positive and negative sequence harmonic). In this set of experiments the raw harmonic data was applied to the PNN network. The orientation angle was very similar to the raw phase angle (Table 5.7) as the phase angle, ϕ is very small. It was considered unnecessary to adjust the complex data for the phase component as the raw phase and orientation values were virtually the same for the low order harmonics.

The main concern with the PNN technique is to select representative target gestures. In this 'pointing' experiment it is a relative easy task to recognize gestures as the first harmonic orientation clearly indicates the gesture and therefore sophisticated techniques are not necessary. To prove that the PNN could recognize gestures, one of the gesturer's responses (Gesturer B) was arbitrarily taken as the target gesture, using a 12 harmonics vector for all gestures. The results gave a misclassification of 6 gestures from a total of 25 i.e. a. 24% error. These misclassifications occurred with the two gestures that were very similar.

It was recognized that gesturer's B target gestures may not have been the most appropriate target gesture. To calculate a more representative target gesture, with this sparse data, the average of all six gestures was calculated. The results are shown in Tables 6.1 and 6.2 which also show the result of having just 1 or 12 harmonics with this data. The 'spread' variable was set at 1 and changes to this value had no effect on the network's performance.

Gesture	Gesturer A	Gesturer B	Gesturer C	Gesturer D	Gesturer E	Gesturer F
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	**5/3**	3	3	**5/3**	3	3
4	4	4	4	4	4	4
5	**3/5**	3	5	5	**3/5**	**3/5**

Table 6.1 Target gestures made from the average of all 6 gestures, misclassification shown bold, near miss shown after '/'. Harmonic pair = 1.

Gesture	Gesturer A	Gesturer B	Gesturer C	Gesturer D	Gesturer E	Gesturer F
1	1	1	1	1	1	1
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	**3/5**	**3/5**	5	5	**3/5**	**3/5**

Table 6.2 Target gestures made from the average of all 6 gestures, misclassification shown bold, **near miss shown after '/'. Harmonic pair = 12.**

Recognition was improved and misclassification was recorded at 17% (6 misclassified from 30) when using just one harmonic. But further investigation showed that all errors occurred between Gestures 3 and 5, the two gestures that are similar as shown in Figure 5.23. In order to gain a measure of the closeness of these two gestures, the mean squared error difference between the input and target vectors for all gestures of gesturer E was calculated. The errors calculated were 0.08 and 0.07 for gestures 3 and 5 respectively, showing the closeness of the error and the resolving difficulty of the network. Consideration of the first orientation angle for gestures 3 and 5 show angles of 93° and 95° respectively that confirms the similarity of the first harmonic orientation angle and hence complex data values. Increasing the number of harmonics to 12 improved the classification of the network, as shown in Table 6.2, to just 13% (4 misclassified from 30) errors.

The results conclude, as expected, that the better the representative target gesture, the better discrimination between gestures can take place. In addition more harmonics

help the discrimination especially when the first harmonic orientation angles are very similar. The choice of target gesture is discussed in more detail in the next chapter using clustering techniques. However, classification and recognition of the pointing gesture experiment is first conducted with clustering techniques to ascertain a rationale for the selection of distance metric and linkage method from results that are well understood.

6.5. Recognition using Clustering Techniques

The hierarchical clustering technique was used because it enables the exploration of possible clustering sets on relatively sparse amounts of data. The hierarchical technique allows for a variety of distance metrics and linkage methods to be used to explore clustering of the gesture data. Three distance metric were investigated, i.e. Euclidean, City Block and Mahalanobis; and up to five linkage methods, were applicable for the distance metric, i.e. single, complete, average, centroid and ward. Before applying these techniques to the data consideration of normalisation of the data is necessary.

6.5.1. Testing clustering techniques

The investigation of clustering techniques using a variety of distance metrics and linkage methods was undertaken under the different normalisation regimes. In addition the validation of the methods described by Jain et al. (1999) were investigated, namely an *internal* examination of validity using the 'Cophenetic Coefficient' and an *external* assessment of validity compares the recovered structure with the prior classification of the gestures. For these tests the pointing gesture experiments were used.

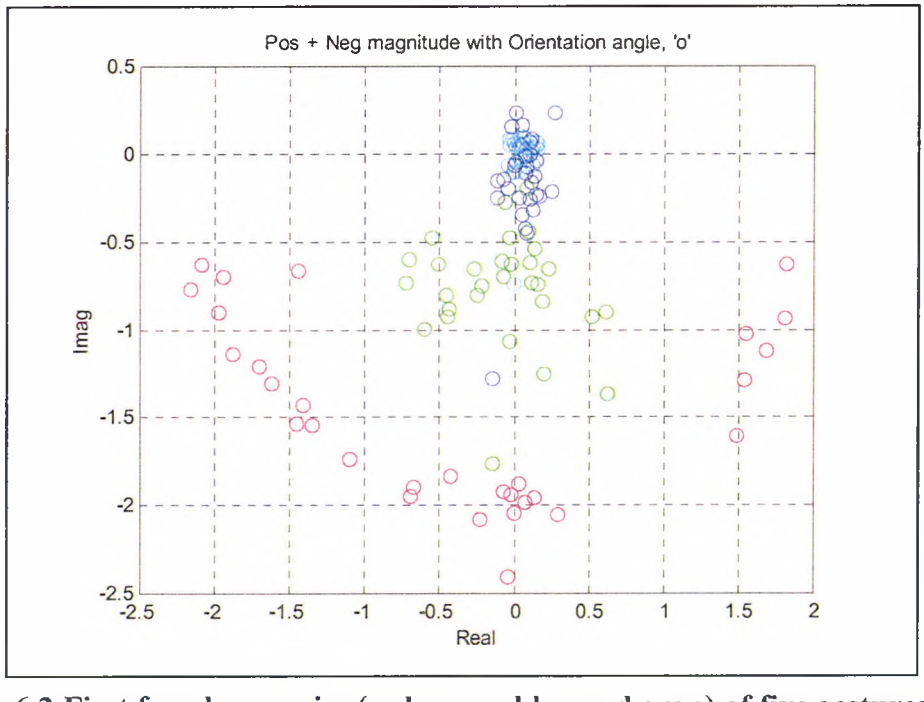


Figure 6.2 First four harmonics (red, green blue and cyan) of five gestures from six gesturers.

The distribution of the first four harmonics for the five gestures of six gesturers is shown in Figure 6.2. The data for Figure 6.2 was obtained from arbitrarily normalising data to the first positive sequence harmonic magnitude.

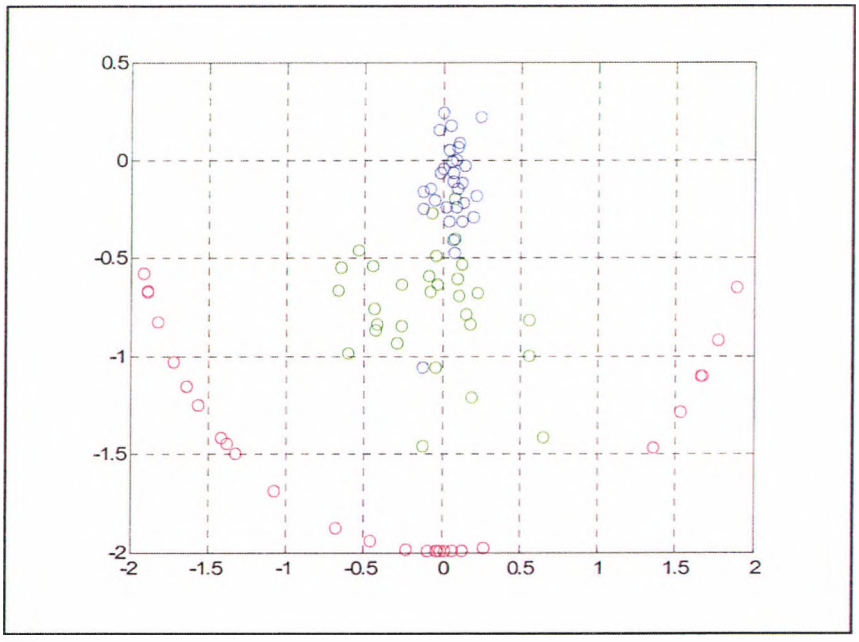


Figure 6.3 First four harmonic (red, green blue and cyan) of five gestures from six gesturers – normalised to magnitude of two for first harmonic

Harmonic	Positive Magnitude, A_p	Negative Magnitude, A_n	Relative Orientation θ
1	0.8952	1.0000	-133.5720
2	0.4637	0.4583	-116.3818
3	0.0668	0.0935	-44.8433
4	0.0450	0.0928	25.3509

Table 6.3 A_p, A_n and Orientation angle for the first 4 harmonics normalised by A_n, for the gesture A1.

Harmonic	Real Magnitude	Imaginary
1	-1.3785	- 1.4490i
2	-0.4323	- 0.8716i
3	0.1199	0.1193i
4	0.1314	+ 0.0623i

Table 6.4 Values of Table 6.3 normalised to the first harmonic magnitude of 2 showing real and imaginary coordinates.

Applying the normalisation technique previously discussed gives the results shown in Tables 6.3 and 6.4. These results are for gesture A1 and show that the first negative sequence harmonic is larger than the positive sequence component. The latter table gives an example of the coordinate values when normalisation to 2

occurs. This new normalisation of the data shown in figure 6.2 results in figure 6.3. The first harmonics (red circles) are clearly shown to be at a radius of 2 in the figure.

Visual interpretation by orientation angle has been discussed in Chapter 5 and the first harmonic orientation angles shows three distinctive conditions and two overlapping conditions as already established by the PNN classification technique. However, experiments were performed to establish how well clustering techniques were able to group the data using the Matlab Statistics Toolbox as detailed in Appendix VI. The 'cophenetic correlation coefficient' was used to determine the best combination of distance measure and linkage method, with the coefficient closest to one being the best solution. The experiments were conducted for the three distance measures of 'City Block', 'Euclid' and 'Mahalanobis', and all five possible linkage methods of 'single', 'complete', 'average', 'centroid', and 'ward' were appropriate as the Mahalanobis is only meaningful using single, complete and average linkage techniques.

The results of the investigation, when using the coordinate values representing the magnitude and orientation of the first harmonic orientation angle are shown in Table 6.6. The 'ward' linkage and the 'Euclidean and City Block distance measure have the highest and most similar cophenetic correlation coefficient. The commonly used Euclidean and Mahalanobis distance measures with the 'single' (nearest neighbour) method do not perform that well in comparison.

Linkage/ Distance	**'single'**	**'complete'**	**'average'**	**'centroid'**	**'ward'**
City Block	0.458	0.631	0.607	0.600	0.639
Euclidean	0.528	0.630	0.619	0.615	0.647
Mahalanobis	0.413	0.546	0.536	NA	NA

Table 6.5 Cophenetic correlation coefficient values when comparing of distance metrics and linkage methods for the first harmonic vector, for normalisation arbitrarily chosen as A_p=1

Linkage/Dist ance	**'single'**	**'complete'**	**'average'**	**'centroid'**	**'ward'**
City Block	0.782	0.880	0.889	0.887	0.835
Euclidean	0.814	0.872	0.874	0.876	0.784
Mahalanobis	0.719	0.886	0.867	NA	NA

Table 6.6 Cophenetic correlation coefficient values when comparing of distance metrics and linkage methods for the first harmonic vector, for normalisation when A_p or A_n is the greatest

Linkage/ Distance	**'single'**	**'complete'**	**'average'**	**'centroid'**	**'ward'**
City Block	0.782	0.877	0.889	0.886	0.838
Euclidean	0.810	0.871	0.873	0.873	0.788
Mahalanobis	0.739	0.885	0.876	NA	NA

Table 6.7 Cophenetic correlation coefficient values when comparing of distance metrics and linkage methods for the first harmonic vector, for normalisation of the first harmonic to value 2.

The output value, c, is the cophenetic correlation coefficient. The magnitude of this value should be very close to 1 for a high-quality solution. This measure can be used to compare alternative cluster solutions obtained using different algorithms.

Data of vector coordinates from the 6 gestures and 5 gestures (30 coordinates) were subjected to clustering using different distance metrics and linkage methods. The results in Tables 6.5, 6.6 and 6.7 show the value of the cophenetic correlation coefficient using the three possible normalisation techniques. Table 6.5, gives the value of the cophenetic correlation coefficient when normalisation is arbitrarily calculated with the value of A_p. The next two tables show significantly higher values as they are a result of normalisation occurring due to the highest value of A_p or A_n.

The cophenetic correlation coefficient is just one method of validating a clustering technique and it shows that arbitrary normalisation does produce overall lower values. It also shows that there is not a great deal of difference between distance metric and linkage method in Table 6.6 or 6.7. The greatest difference in value is shown with the single linkage method and the Mahalanobis and Euclidean distance metric, which differ by just 7%.

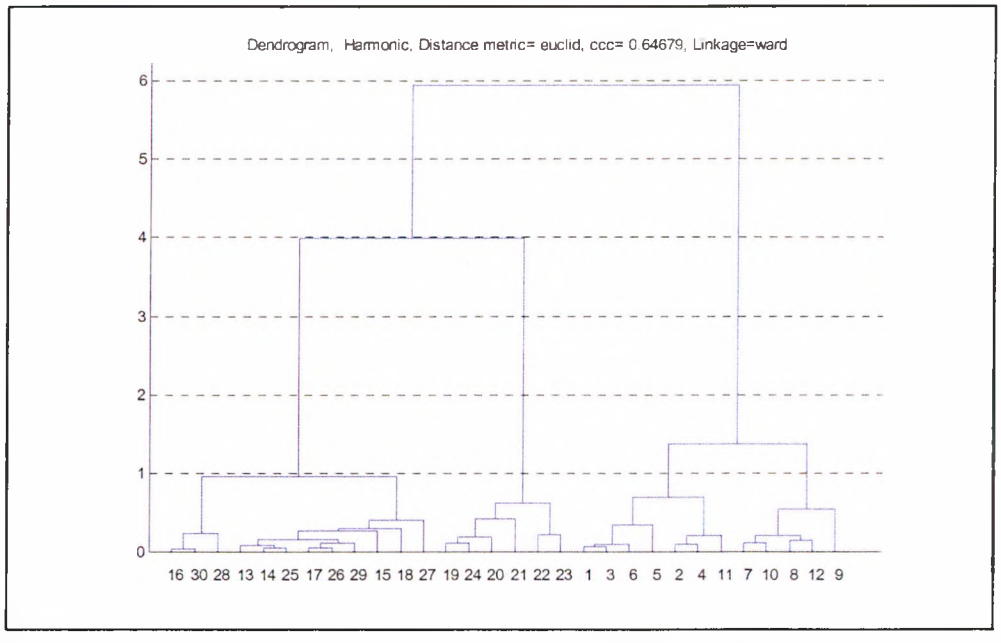


Figure 6.4 Dendrogram of 30 gestures using Euclidean distance metric and 'ward' linkage method for the first harmonic vector.

Another way to validate clustering is by comparing the dendrogram of all the experiments produced in the previous experiments. A range of dendrograms for various distance metrics and linkage methods are to be found in Appendix VI. They establish the clustering forms described by Jain et al.(1999), in that the single-link algorithm suffers from a chaining effect and has a tendency to produce clusters that are straggly or elongated. The dendrogram of the Euclidean solution using the 'ward' linkage is shown in Figure 6.4, and shows particularly clear groupings. These groupings appear to be very similar to those that a human observer would choose and

also give results based on the objective prior knowledge of the data. The shorter links show consistency but the longer links show inconsistency where natural clustering occurs in the data. The numbers on the 'x' axis of the dendrogram relate to the total number of gesture instances, there being thirty in total. For example, the numbers 1 to 6 relate to the results of six gesturers (A, B, C, D, E & F) to the first gesture, and so on. Gesturer A's response to the five gestures is shown as gestures 1, 7, 13, 19 and 25.

Gesture	Gesturer A	Gesturer B	Gesturer C	Gesturer D	Gesturer E	Gesturer F
1	3	3	3	3	3	3
2	4	4	4	4	**3**	4
3	1	1	1	1	**2**	1
4	5	5	5	5	5	5
5	**1**	**1**	1	**1**	2	2

Table 6.8 Classification of gestures for each gesturer using 'euclid' distance metric and 'ward' linkage method (bold shows miss-classification).

Gesture	Gesturer A	Gesturer B	Gesturer C	Gesturer D	Gesturer E	Gesturer F
1	3	3	3	3	3	3
2	**3**	**3**	**3**	4	**3**	**3**
3	1	1	1	1	1	1
4	5	5	5	5	5	5
5	**1**	**1**	**1**	2	**1**	**1**

Table 6.9 Classification of gestures for each gesturer using 'euclid' distance metric and 'single' linkage method (bold shows miss-classification).

The classification experiment was repeated, but instead of using the PNN as before, the coordinate values from the vector that represents the magnitude and phase of the first orientation angle was used. The classification results using the Euclidean distance measure with 'ward' and 'single' linkage are shown in Tables 6.8 and 6.9 respectively. The Euclidean distance metric performs well but is influenced by the linkage method as the misclassification drops from six to ten when the 'single' method is used instead of the 'ward' method. As a result the Mahalanobis metric did not improve results but gave poorer clustering when required to cluster data into five groups.

The clustering results were compared with those obtained with the PNN network. Table 6.10 shows the different allocation of gesture number that occurred for the PNN and Clustering methods. The gesture numbers of the clustering technique were aligned to the numbers used in the PNN experiments, as shown in Table 6.10. A comparison of the clustering using the PNN method and the Clustering method is shown in Table 6.10. There are similar results from both methods. The PNN method has 5 from 30 wrong classifications. The cluster method has 5 or 8 misclassifications as there is no way of knowing what gesture is correctly labelled for gestures 3 and 5 and which is the correct allocation, whereas the PNN definitely misclassified.

Action of right hand	Gesture PNN	Gesture Cluster
To left shoulder and return	1	3
To the left and return	2	4
Straight up and down	3	1
To the right of subject and return	4	5
Straight ahead, 'halt' and return	5	2

Table 6.10 Gesture Number and Action Alignment

Gesture	Gesturer A		Gesturer B		Gesturer C		Gesturer D		Gesturer E		Gesturer F	
	P	C	P	C	P	C	P	C	P	C	P	C
1	1	1	1	1	1	1	1	1	1	1	1	1
2	2	2	2	2	2	2	2	2	2	**3**	2	**4**
3	**5/3**	3	3	3	3	3	**5/3**	3	3	**2**	3	3
4	4	4	4	4	4	4	4	4	4	4	4	4
5	**3/5**	**3**	3	**3**	5	**3**	5	**3**	**3/5**	**5**	**3/5**	5

Table 6.11 Comparison of PNN method (P) and Clustering method (C) for classification using one harmonic, (bold shows miss-classification).

6.5.2. Alternative clustering input data

The previous work using the PNN method clarified that using more harmonics improved the classification rate, albeit slightly. However, a vector based on the first orientation angle appeared to be an important contributor to gesture classification. An alternative approach for data input is to use just orientation angles instead of the coordinate values. The input must have a minimum dimension of two so the first and second orientation angle was input for one experiment and the first three orientation angle was in put for another experiment.

Using just orientation angles gave generally worse results than using coordinate data of the first harmonic vector. When using two orientation angles, only one gesture was classified correctly; the misclassification was 20% (6 from 30). When three orientation angles were used the classification slightly improved to 30% (9 from 30). The poor performance is attributable to several factors. Firstly, there could be a lack of scaling of the second and third harmonic components that could have an undue influence on the result. The problem to establish is what scaling should be used on the orientation angles, as the angle does not have any amplitude component as with a vestor. Secondly, the clustering technique may be differentiating between gestures by taking into account the second and/or third orientation angle that characterise a gesture.

It becomes apparent that the PNN technique is better than the clustering technique because of the automatic scaling that is inherent in the PNN network. However, the main challenge with the PNN technique is to find appropriate target gestures to represent a gesture from a group of people. The first harmonic vector does characterise the overall gesture but the second and third harmonic seem to generate

characteristic of an individual gesturer or a group of gesturers. The possibility of using the clustering technique to select the target vectors using the second and third harmonic information is considered in the next chapter.

6.6. Summary

The recognition of gestures from trajectory data is a case of temporal pattern recognition. In this chapter the use of the PNN and the use of hierarchical clustering is explained. The PNN, based on the RBF, may require more neurons than standard feed-forward back-propagation networks, but can be designed and trained in a fraction of the time it takes to train standard feed-forward networks. The PNN requires no training and is very easy to apply. The key factor in using PNN is to obtain appropriate target gestures.

It has already been established that the harmonic orientation angles from the trajectory data characterise the gesture. It is shown that the first orientation angle is adequate to classify a range of pointing gestures from different people. When the harmonic data obtained from one of the gestures is used as the target gesture in a PNN network the other gestures are readily classified. Improvements can be made to recognition rates by introducing more harmonics to the analysis; however the improvements were small in the pointing gestures. The cause of misclassification was mainly due to the very similar characteristics of the two gestures.

An alternative technique for classifying data is to use a clustering technique. Hierarchical clustering techniques were favoured over the k-means technique because it was ideal for small amounts of data. It also readily allowed experimentation with distance metrics and linkage methods to ascertain optimum clusters based on a number of validity tests. From these tests two important results were obtained. One result clarified the type of normalisation that should be applied to the data. It was discovered that normalisation of the frequency data by the greatest of the positive and negative sequence component was relevant and showed a higher correlation with cophenetic distances. It also solved some problems with some experimental results in the next chapter.

The data fed into the clustering function represented the real and imaginary coordinates of a vector representing the magnitude (A_p+A_n) and orientation of the first orientation angle. Adjustments to the distance metric and the linkage method showed that for this data the Euclidean or City Block metric and the 'ward' (incremental sum of squares) linkage method gave results comparable to those of the PNN. But with this data set the Mahalanobis distance metric did not perform better than the other distance metrics.

The difference between the classification techniques i.e. just using the clustering tool or using the PNN, showed some subtle differences. For example, with the two overlapping different gestures, using the clustering technique, it was not possible to determine if the class was the correct. However, with the PNN network it was clear if the classification was correct or not. A further complication results if the dimensions of the clustering data are extended from their minimum of two (typically coordinate data). The second harmonic would be typically at lower amplitude than the first harmonic. This difference in scale can affect the proximity calculation and produce

an erroneous result. Although scaling could be applied there is a problem of adjusting it for an optimum performance. The PNN does not suffer from the scaling problem, as the comparison of the input with the target gesture is automatically performed in the radial basis layer.

The main concern with the PNN technique for classification is determining the optimum target gesture for classification. The clustering technique can be used as a tool to extract the optimum target vector. It appears that the first harmonic orientation angle can be an automatic clustered value due to spatial constraints of the gesture, and with the next few harmonics (second and third) can be a vehicle to characterise a gesture.

The next chapter will investigate the use of harmonics in both the classification and recognition of gestures to establish how they can be used to detect intra-class and inter-class differences in a gesture.

7. Gesture Experiments

The aim of this chapter is to assess the performance of Fourier analysis, clustering and PNNs techniques on different types of single-handed gesture. Experiments were performed on a hand-raising gesture that was repeated ten times to assess the intra-class performance. Another single-handed gesture experiment was performed with twenty-one people to assess the inter-class performance. This gesture was called the 'Take-Mug' experiment as the gesturers mimicked taking an imaginary cup and drinking from it. The route of the trajectory from these experiments was much more complicated than any used before.

The Fourier analysis of these trajectories required the full explanation of the properties of the significant harmonics. Observation of the individual gesturer's actions suggested that within the gesture there were subtle variations that could be additionally grouped or clustered together. Clustering methods were employed to resolve target gestures that could be used with a PNN. The clustering inputs were based on vectors made from the amplitudes of the positive and negative sequence components and the orientation angles of the first three harmonics. The results from the application of clustering and PNNs were compared with visual observations. A further set of single handed gestures were performed with a number of people to investigate further mime like type of gestures often found in communication interactions of normal and handicapped people. These gestures gave some interesting results of which the Fourier analysis technique was an ideal tool to reveal the oscillatory nature of many of these gestures. These gestures also revealed limitations of the tracking and analysis technique.

7.1. Introduction

Previous gesture experiments tracked hand movement by the combination of motion and skin-colour cues to form gesture objects, which are sorted into rank order by area. The basis of this hypothesis is that in a situation with a single person in front of a camera, and no other people moving in the background, there were likely to be only three skin-coloured objects possible in a gesturing image sequence (the head, and two hands). The study of single-handed gestures showed that the trajectory of the hand could be recorded in the spatial domain and transformed into the frequency domain by the FFT. The recognition of gestures could take place using the harmonic components in a manner similar to the way in which Fourier Descriptor technique is used to recognise objects.

However, unlike the Fourier Descriptor technique that is based on relative magnitude of harmonics, it was found that the phase components associated with each harmonic held additional important information about the characteristic of the gesture. The highly dimensional harmonic data (real and imaginary component of the positive and negative component) lent it self to be easily used with a PNN. The use of clustering techniques to identify target gestures for the PNN allowed classification and recognition of gestures to take place.

Additionally, it was found that only the first few harmonics were sufficient to characterise the trajectory. This approach was substantiated by taking the IFFT of a

few low order harmonics that showed that the resultant time domain waveform closely matched the original waveform or trajectory.

The frequency components showed that there was a particular structure to the gesture. The structure uncovered by this technique complements the structure and characterisation discussed by Rossini (2003) and Gibet et al. (2001), relating to position of gesturing and shape of gesture. Further experiments were devised to test the system for it characterisation, categorisation and recognition. The first set of experiments was directed at deictic or pointing gestures (Chapter 6) similar to the work of Howell et al. (1998). Experiments were conducted into the publicly available PETS sequences of the same gesture (arm being raised several times) and were the foundation for an investigation into intra-class variations.

A more complicated gesture was inspired by an avatar (obtained from Kennaway, 2004) with twenty-one people imitating the gesture. This gesture was to pick up and drink from an imaginary mug. The avatar was a convenient form of a repeatable, non-varying gesture for the people to be acquainted with. Although each gesture was unique and the key gesture action was observable, inter-class variations could also be extracted. Finally, a series of gestures were obtained from a group of people as a result of showing them a number of stimuli. In order that they were not influenced by observations of others or the experimenter, each person was separately shown a drawing of an object and asked 'What would you do with this?' A drawing of an object, like a hairbrush, saw or toothbrush was shown. Woll (2004) suggested these types of gesture stimuli as it is often used in the treatment of disabled people with communications difficulties. The communication with a basic gesture is often a precursor to any possible sign language learning that the disabled people might be able to undertake at some later stage in their training. Of course non verbal communication is also undertaken by 'normal people' (Chapter 1), but often overlooked in communications because of the difficulty of recording the actions along with spoken or written word. It was already known that even the most experienced therapist sometimes had difficulty recognising some disabled people's gesturing (RNIB, 2004). It was considered useful therefore to see how normal people reacted to these gesture stimuli.

This chapter reviews the findings of these sets of experiments. The sequence of experiments, apart from those publicly available was recorded with either a Logitech web cam or with a Panasonic (NV-DS60B) digital video camera. These recorders were used with both automatic light level control and white balance control which were inconsistent. This gave some challenging image sequences to work with.

The experiments gave some unexpected results that challenged the basis of the experiments. This instigated reflection on the nature of gesture and the categorisation that had been mentioned by other researchers. The tri-phasic gesture categorises the gesture into the transition from rest; the stroke in gesture space and then the transition to the rest state again but is only relevant to the simplest gesture. The 'Take Mug' gesture showed that the stroke phase can have multiple components and varies from person to person. In addition, depending on the gesture, the instructions and the people involved, different interpretation and reactions to the stimuli were observed. The experiments with the gesture stimuli were anticipated to be simple and easily recognisable. However, the gestures were found to be far more complicated and perhaps even more complicated than sign language, which although appearing

complex to a non-signer has been formalised and learnt systematically so that it is very repeatable with variations at a minimum between signers.

7.2. Repeated gesture (Intra class variations)

A gesturing sequence was selected to investigate intra-class variations. The gesture selected was a repeated hand raising gesture (Appendix II and VII) from the PETS sequence of data discussed in Chapter 4. The sequence comprised of ten gestures of raising the right hand. The row and column coordinates of the gestures are shown in Figure 7.1 Recording at 25 frames per second gave gesture lengths varying between 32 and 42 frames with an average of 37.

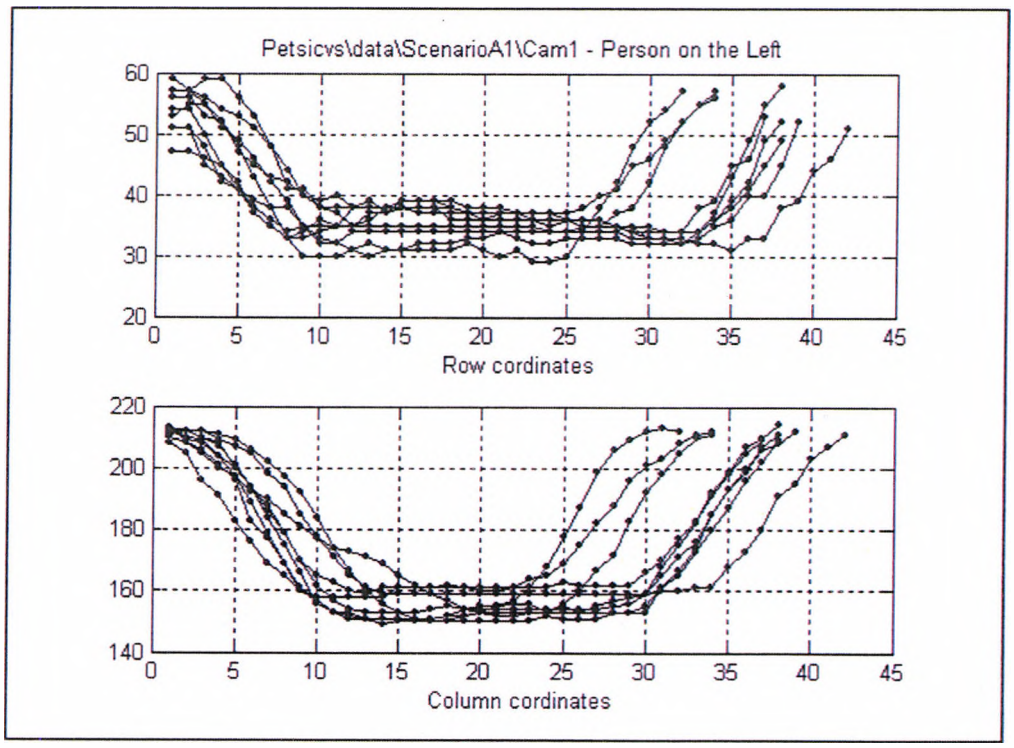


Figure 7.1 The row and column coordinates for the ten repeated hand raising gestures.

The harmonic analysis is shown in the Appendix VII and shows the tight clustering of the first orientation angle. Figures 7.2 and 7.3 shows the tight clustering of the first harmonic whether normalised by A_p or normalised to the vector magnitude of 2, respectively as explained in the previous chapter.

It was shown in Chapter 5 that the first orientation angle related to the spatial angle subtended between the start of the gesture and the high point of the gesture for the pseudo triangular gesture. Data was recorded both manually/visually and by the automatic method and gave very similar average values of -14° and -9° respectively for the orientation angle, showing that the hand start/stop location and high point remained very consistent for all ten gestures. The standard deviation was recorded at 5° and 6° respectively, virtually the same and consistent with experimental error. It is noticed that the second harmonic also remains quite tightly clustered, averaging at -14° with a standard deviation starting to increase at 9°.

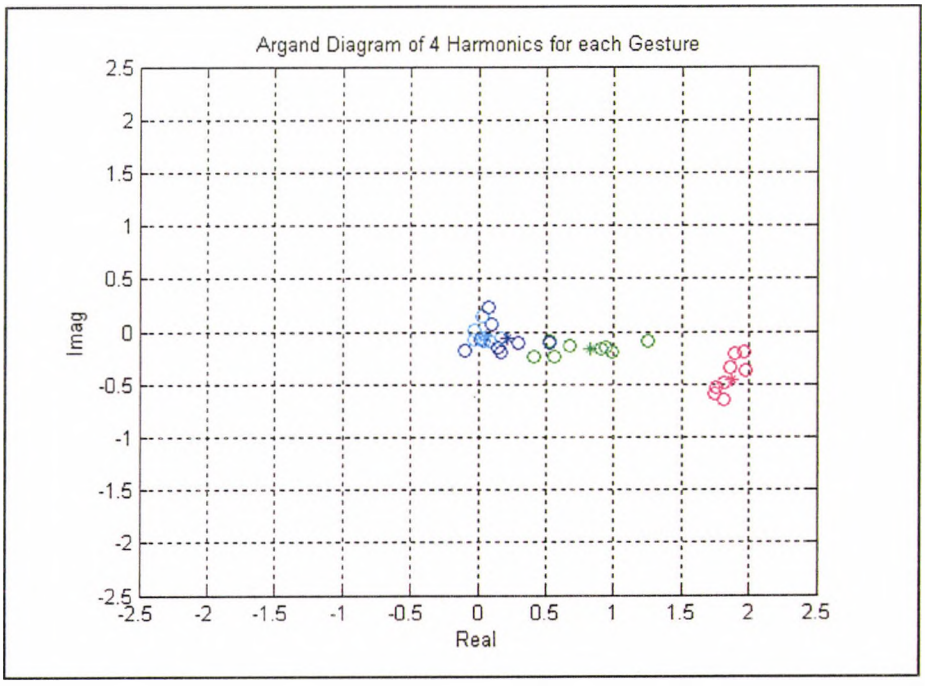


Figure 7.2 The first four harmonic orientation angles for ten repeated hand-raising gestures. Harmonics 1-4 are red, green, blue and cyan, shown as 'o' respectively, and average value shown as '*'. Vector magnitude normalised by A_p

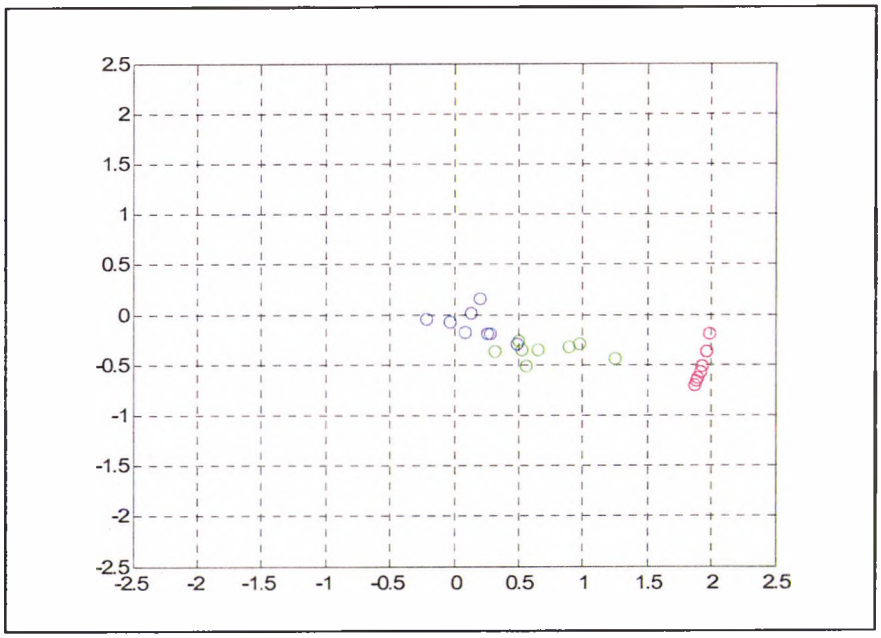


Figure 7.3 The first four harmonic orientation angles for ten repeated hand-raising gestures. Harmonics 1-4 are red, green, blue and cyan, shown as 'o' respectively, and average value shown as '*'. Vector magnitude normalised to 2.

The main differences in harmonic content occur with the third harmonic were the average values are similar at -22° and -12° respectively, but the standard deviation values are quite large at 54° and 19°, although the magnitudes have reduced to about a fifth of the first harmonic. However, as explained in the previous chapter the standard deviation and mean calculation can be unreliable as angles become more scattered because of the discontinuity problem in angle measurements. It is noticeable that the magnitude of the third harmonic had a considerable variation in magnitude between gestures.

The similarity of the first orientation angle shows that the ten gestures are all the same. However, individual idiosyncrasies are revealed by the other harmonics, mainly by the second and third harmonics because they usually have much larger amplitudes than higher order harmonics. There is a difference between the shortest (gesture 5) and longest (gesture 6) gesture. Observation of the image sequence seems to suggest that the hand in the longest sequence remains at the top of the trajectory, statically and longer than the shorter gesture. The frequency response data shows the difference between the two gestures relates to the difference in amplitudes of the second harmonic orientation angles, although the orientation angles are similar. This would suggest that there is more curvature in the longer trajectory. The effect of the second and third harmonics is considered in more detail in the following section regarding the 'take-mug' experiments.

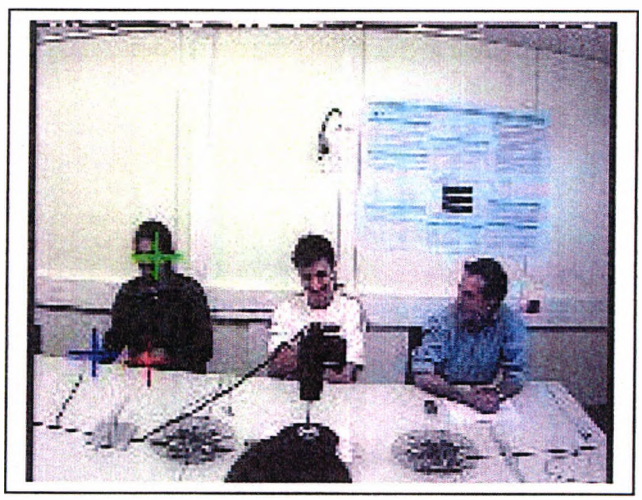


Figure 7.4 A JPEG image showing the right hand (blue '+', 3rd SCM object) about to disappear at the end of the sequence. The left hand (red '+', 1st SCM object) rising for 10 frames. The head is also detected moving (green cross, 2nd SCM object)

Before considering the role of the low order harmonics there is one additional observation to be made regarding the repeated hand gestures. At the beginning of one of the sequences the gesturer was laughing and at the end of the sequence the left hand appears to be raised involuntarily for ten frames. The movement of the left hand could not be classified as any substantial trajectory but more like a computer 'flag' being initiated. It is also worth noting that the head made a significant movement at the end of the right-hand trajectory. Movement of both hands and head are detected, as shown in Figure 7.4. The sudden movement of the left hand is indicated by the

first SCM object (red), followed by the right hand be assigned to the second SCM object (green) and finally the lesser movement of the head is shown by the third SCM object (blue) location as detected by the first three SCM objects

The main conclusion from this experiment is that the first harmonic vector, comprising of the normalised amplitude and orientation angle of the first harmonic, were closely clustered together. This was a similar result to those gained with the 'pointing gestures' in chapter 6. However, the variations in the gesture profile in shape and duration was reflected in mainly the properties of second and third harmonic components. As a consequence of these observations a more complicated gesture trajectory was investigated that showed larger variations in the second and third harmonic properties. These experiments are described in the 'Take-Mug' experiment of the next section.

7.3. The 'Take Mug' Gesture (Inter class variations)

The 'Take-Mug' experiment was devised to show that the recognition method could discriminate between subtly different versions of the same gesture. The experiment consisted of recording image sequences of twenty-one people mimicking the action of taking a mug to drink; lifting it; drinking and then putting the mug down as inspired by the avatar animation (Appendix I). A typical gesturer is shown in Figure7.5, and from each sequence coordinate data was manually produced for subsequent analysis. An example of the image sequence of the gesture shown in Figure 7.5 is shown in Appendix VII.

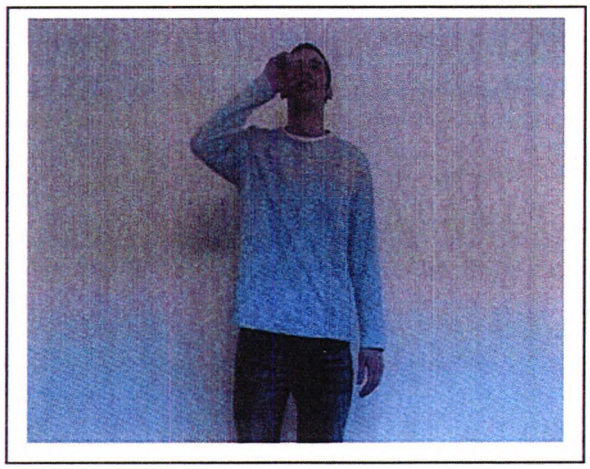


Figure 7.5 An image from a 'Take Mug' sequence

Data from this type of gesture was found to last longer than the pointing type of gesture. It was also found that this type of gesture generated a wider range of harmonic characteristics than with previous pointing gestures. This resulted in more careful consideration of the normalisation procedure both in terms of the time normalisation procedure and the vector normalisation. In addition, the gestures of the twenty-one gesturers were seen to fall into distinctive sub-classes of the basic 'Take Mug' gesture and were confirmed by clustering techniques using just the second and third harmonics.

7.3.1. Sampling Rate Adjustments

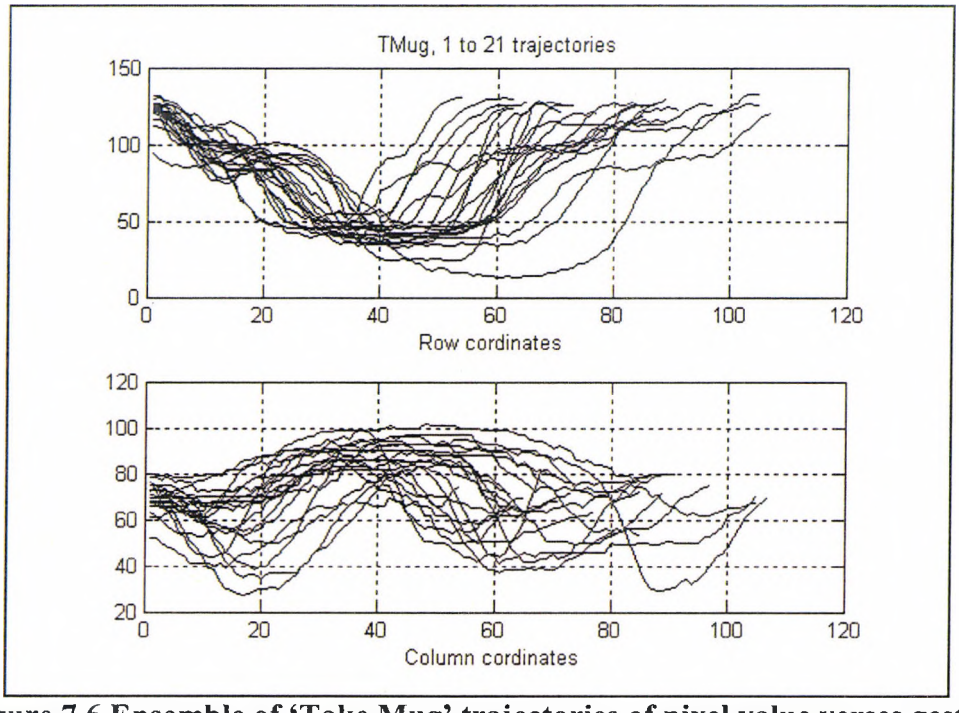


Figure 7.6 Ensemble of 'Take Mug' trajectories of pixel value verses gesture length for row and column coordinate data.

The length of the twenty-one 'Take Mug' gestures ranged from 57 to 103 frames. The row and column coordinates are shown in Figure 7.6. In previous chapters it had been discussed that gesture length would be normalised to 64 samples, which clearly could not be realised for sequences over 64. To compensate for these longer sequences several strategies were considered. The normalisation length could be changed to say 128, or the LUT could incorporate a compression factor for gesture length greater than 64. The technique implemented was to skip every other sample, so as reduce the sequence length by half. The frequency response of a sequence originally of 63 samples is shown in Table 7.1. The new response, when the gesture length is reduced to 32 is shown in Table 7.2.

Harmonic	Positive Magnitude	Positive ϕ	Negative Magnitude	Negative ϕ	Orientation θ
1	1	38	1.08	-38	23
2	0.31	26	0.25	-26	-17
3	0.35	-5	0.27	5	10
4	0.16	18	0.13	-18	-18
5	0.08	-25	0.02	25	2
6	0.05	-64	0.01	64	15
7	0.04	-100	0.01	100	26

Table 7.1 Frequency Response of Sequence A of 'Take Mug' suite of Experiments, original length of 63.

Harmonic	Positive Magnitude	Positive ϕ	Negative Magnitude	Negative ϕ	Orientation θ
1	1	33	1.08	-33	22
2	0.33	14	0.25	-14	-16
3	0.36	-17	0.28	17	10
4	0.15	-1	0.12	1	-18
5	0.09	-58	0.03	58	15
6	0.05	-93	0.01	93	11
7	0.05	-113	0.03	113	6

Table 7.2 Frequency Response of Sequence A of 'Take Mug' suite of Experiments, reduced to 32 from original length of 63.

Comparison of the results shows minimal difference between the two sets of data. The first harmonic shows very close similarities in magnitude and phase values. The orientation angles for all harmonics show remarkably similar values. The most significant variation is with the fifth harmonic showing a variation of 2° and 15°. However, the magnitudes of the positive and negative sequence components are very small and so some inaccuracies would be expected. The results show that half sampling rate is sufficient for this set of experiments.

7.3.2. Analysis of the 'Take Mug' Gesture Suite

The analysis of the frequency response shows some interesting results. The first orientation angle for twenty of the twenty-one subject's has a standard deviation of 2° with a mean of 22°, a very tight cluster. In the previous chapter normalisation processes were discussed. It was found that for simple trajectory paths normalisation by the positive sequence component A_p was a quick and easy method of normalisation for scale invariance of the gesture. However, as trajectory paths became complex this technique became unsatisfactory and normalisation to a vector of magnitude of 2 was critical for clustering and comparison techniques to take place. Figure 7.7 and Figure 7.8 show the results of the two normalisation methods on the first six harmonics.

The orientation angle on its own gives a good indication of the clustering characteristics. However, both the PNN and clustering techniques require inputs in the form of coordinates. These coordinates are formed from a vector represented by the amplitude of each harmonic component and the orientation angle. Adding the magnitude of the positive and negative sequence components together forms the amplitude of the harmonic. Formerly, the positive sequence component had been arbitrarily set to unity and all other harmonics scaled accordingly. The orientation does not change whether the positive or negative sequence component is the greatest. However, the magnitude of the vector can vary considerably if the negative sequence component is prominent and this can cause error in the cluster and PNN analysis. This situation is rectified by determining whether the positive or the negative

component is the larger and scaling by the larger component. In so doing an anti-clockwise path or clockwise path is determined.

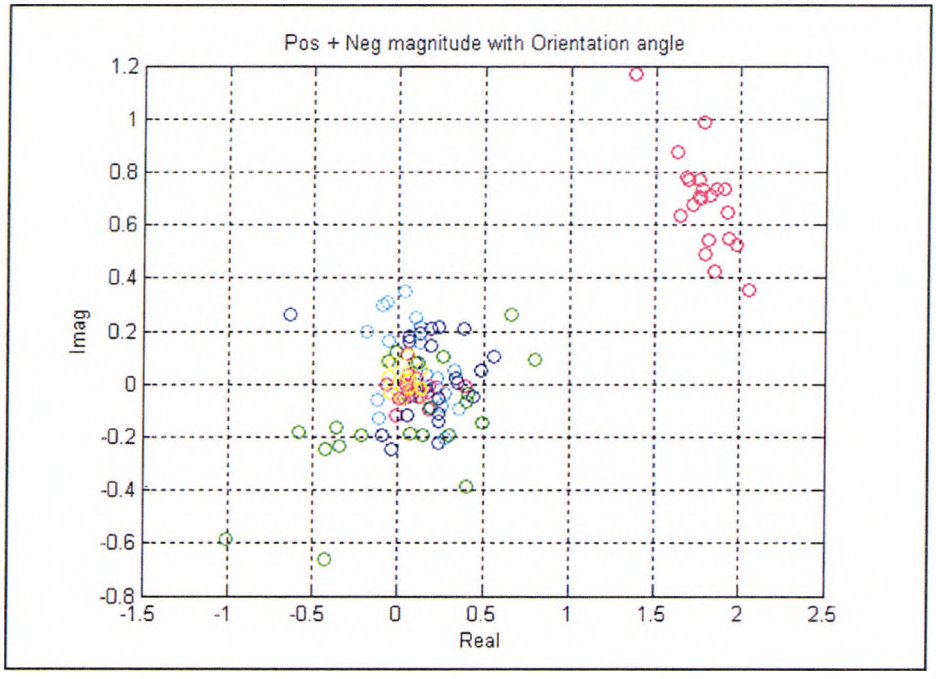


Figure 7.7 First six harmonics vectors (red, green, blue, cyan, magenta and yellow respectively) of twenty-one gesturers performing the 'Take-Mug' gesture, normalised to Ap equal to 1 for the 1st harmonic.

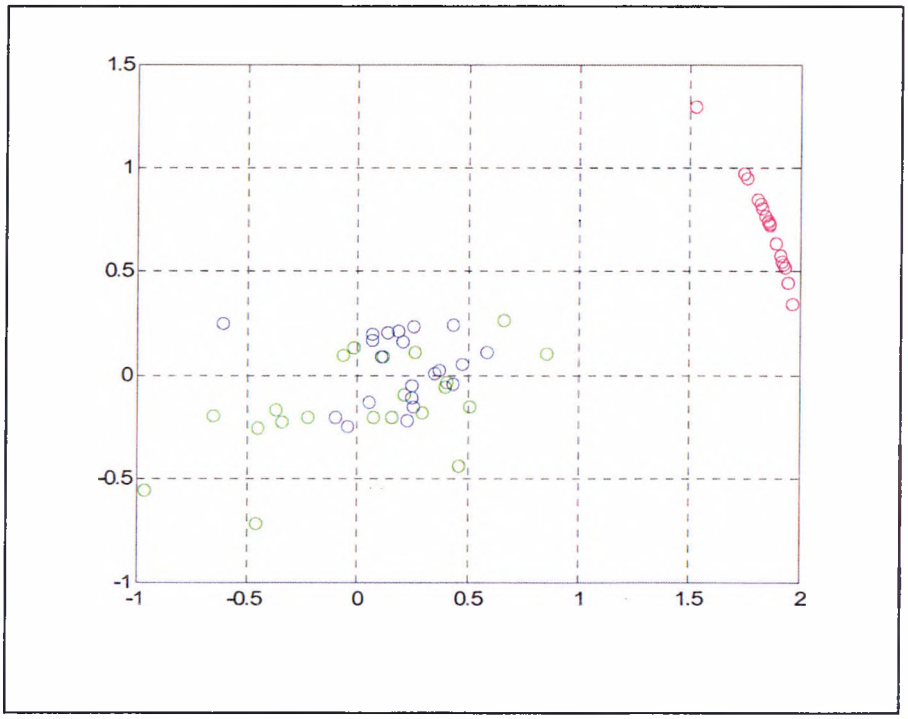


Figure 7.8 First three harmonics vectors (red, green and blue respectively) of twenty-one gesturers performing the 'Take-Mug' gesture. Normalised to Ap plus An of 1st harmonic equal to 2

7.3.3. Variations of the 'Take Mug' Gesture Suite

Although the miming of the 'Take Mug' gesture was fundamentally the same the interpretation produced considerable variability in the 21 sequences, as listed in Appendix VII, but can potentially be divided into different categories. For instance, some gesturers take the imaginary mug just to their right side, move it up to their lips and then return the hand to their starting position. This is similar to the avatar motion but probably less jerky and slower. Variations on this theme are to return the jug, after drinking, to the imaginary table where it was taken and this can sometimes be a different spatial position to where it was picked up from. The taking of the mug can also be in front of the person. It can also be picked up in a position very wide of the individual to the right hand side. Some individuals, take the imaginary mug and do not continue raising the mug, but let the hand fall for a while before raising the hand to the mouth. Finally at the end of the trajectory the hand does not always return to the starting position by a direct route, but the hand takes a curved arc to the right before coming in quite quickly to the starting position.

Although it was clear that every one of the twenty-one individuals was miming a typical drinking activity, it was also apparent that every action was unique. However, it was clear that the activity could be classified into a series of characteristics that some individuals exhibited and could be recognised in general by the second and third harmonic orientation angles. Of the twenty-one gestures, observations seemed to show that there were four distinct characteristics, which are represented by the gestures α, β, γ and δ. Gesture α appeared to be close to the original definition and similar to the avatar, but with a smoother motion. Gesture β appeared to make a deliberate placement of the mug, whereas gesture γ appeared to use different positions for taking the mug and replacing the mug. Finally the gesture δ was very wide with the hand taking the mug a considerable distance from the body.

For the purpose of categorisation of the gestures by a PNN, the target gestures were taken as gestures A, G, K and M for gesture types α, β, γ and δ respectively. The classifications are shown in Table 7.3 along with the first three orientation angles for each gesture. The classification for example gesture B was target type 'α', by the PNN as shown by a '1' in the table. An indication of how close the other targets gestures were in the classifying the unknown gesture was achieved by calculating the least squares difference between the input gesture and each of the target gesture. In the example, the next closest target was target 'γ' and labelled 2 in the table, and so on. The ranking of target classification was used later to give an indication of the appropriateness of the classification accuracy with an automatic classification technique.

Examination of the clustering of the gesture sub-groupings show a grouping of similar second and third harmonic orientation angles for a particular gesture type. For instance, those gestures categorised as having an 'α' type gesture are seen to have second harmonic orientation angle of about $-16°$ and a third harmonic orientation angle of about $10°$. Similarities can also be seen between the other gestures, but those of type 'γ' seem to be less well defined. This maybe because there is no natural grouping that can be sensibly shown, because of the unique nature of each gesturer's action. Or the observer just found the classification task overwhelming.

Gesturer	Target α Rank Position	Target β Rank Position	Target γ Rank Position	Target δ Rank Position	1st Harmonic Orientation Angle θ_1	2nd Harmonic Orientation Angle θ_2	3rd Harmonic Orientation Angle θ_3
A	1	3	2	4	22	-16	9
B	1	3	2	4	18	-5	-6
S	1	3	2	4	21	-41	26
G	2	1	3	4	21	21	-99
D	2	1	3	4	24	6	-30
F	3	1	2	4	21	19	-39
K	2	4	1	3	15	-51	40
C	2	3	1	4	21	-69	71
E	2	3	1	4	15	-30	45
H	3	2	1	4	16	123	-115
L	3	2	1	4	12	88	55
O	3	4	1	2	21	-150	55
P	3	2	1	4	25	-26	-63
T	3	2	1	4	23	35	-24
Q	2	3	1	4	15	-9	0.2
J	3	4	1	2	21	-137	-12
N	3	4	1	2	21	-155	67
M	3	4	2	1	28	-122	3
U	3	4	2	1	24	-145	6
V	3	4	2	1	40	-162	37
R	3	4	2	1	9	-149	157

Table 7.3 Classification of the gestures into the four sub-classes, α, β, γ and δ of the 'Take Mug' gesture using a PNN and least squares calculations ranks the result to the nearness of the other targets. The first three orientation angles are compared to the classification.

7.3.4. Clustering of 'Take Mug' Harmonics

7.3.4.1. First harmonic orientation angle cluster

In section 7.2 when intra-class variation was discussed it seemed reasonable to expect the first orientation angle to be similar because it was the same person repeating the same gesture. In this experiment there are twenty-one people repeating the 'Take-Mug' gesture and the first harmonic angle remains for the majority of the gestures, surprisingly tightly clustered considering the variability of human physical features. The first harmonic orientation angle, for the twenty-one gesturers are shown in Table 7.3 and shows that the majority of angles is in the region of 20° and there is just one outlier at an orientation angle of 40°. Calculations give the average orientation angle at 21.3° and a standard deviation of 6.5° a variation of 2% in 360°. In addition, the average of all twenty-one first harmonic vectors gives an average magnitude of 1.366 and standard deviation of 0.061 a variation of about 5%. These values are all within experimental error and quite surprising considering the difference in each gesturer's physical size.

However, applying the clustering technique to the first harmonic data gave some varied results. The results had shown that gesturer's first orientation angle appeared significantly different to all the other results. However, when the clustering technique was applied for two clusters, using the Euclidean distance metric and 'ward' linkage method gave a the result as shown in Figure 7.9, with the data being almost evenly divided into two clusters. The 'outlier' gesture was isolated by the clustering algorithm when the 'single' linkage method, as shown in Figure 7.10, which is what is expected from a visual interpretation of the data.

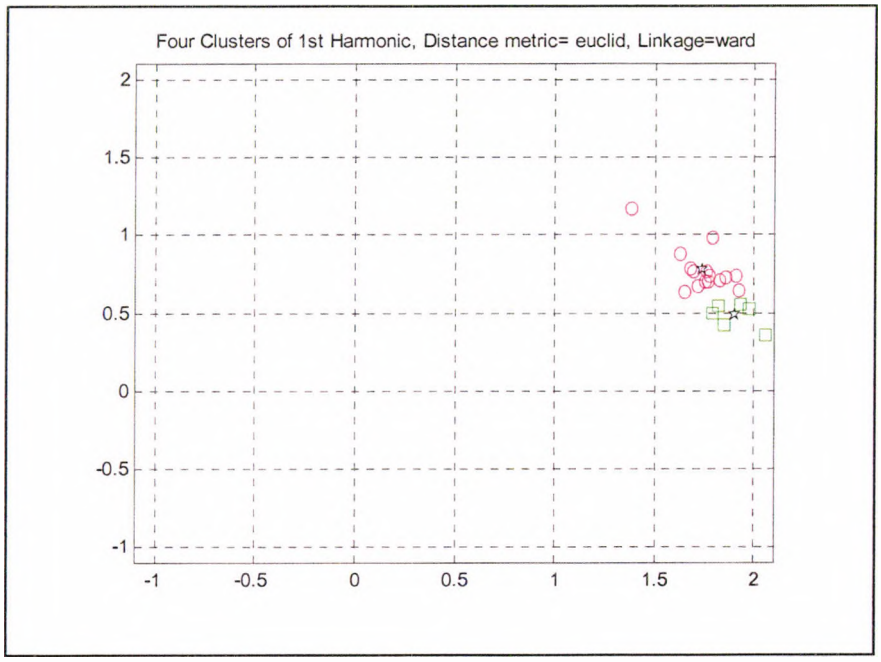


Figure 7.9 Two clusters of the first harmonic using the Euclidean metric and ward linkage method

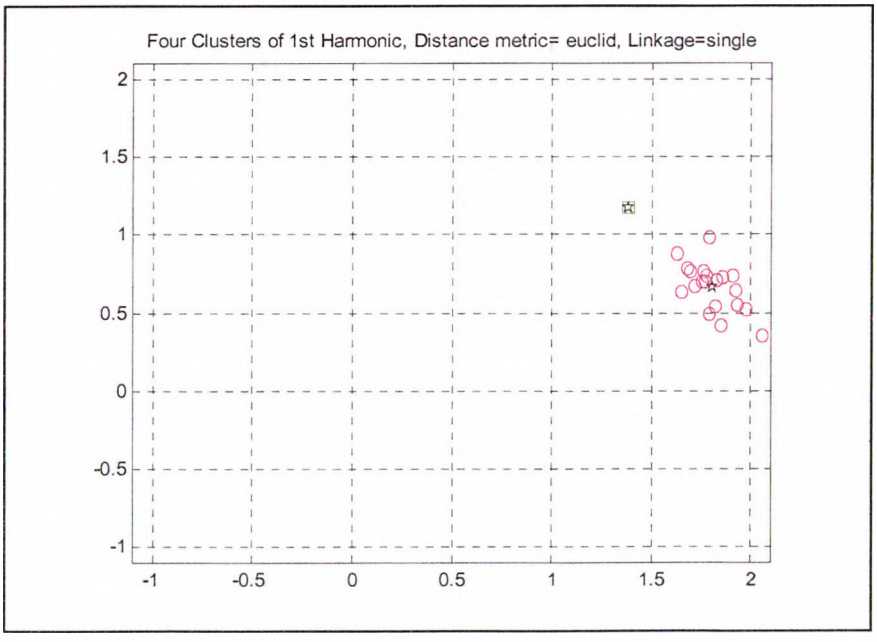


Figure 7.10 Two clusters of the first harmonic using the 'Euclid' distance metric and the 'single' linkage method that isolate the outlier vector.

The use of clustering techniques on data with small deviations in magnitude and angle would appear to be inadvisable. Where the deviation is compatible to experimental error and system resolution, clustering techniques are unreliable. In chapter 5 the resolution of the system was estimated to be about 4°, which is similar in order to the standard deviation of 5° for the first orientation angle. It would appear that the 'single' or nearest neighbour linkage method is more suited to detecting outlier data when the remaining samples are compact. Interestingly, the distance metric and linkage method clearly affect the clustering output. It should be remembered however that as stated by Jain et al. (1999) in chapter 6 clustering algorithms will, when presented with data, produce clusters regardless of whether the data contains clusters or not. The coordinates of the average value of the first harmonic vector were calculated to be 1.2701 and 0.5029 or as a vector, of magnitude 1.366 at an angle of 21.6degrees. These values were used in subsequent experiments with the PNN.

7.3.4.2. Second and third harmonic clusters

Clustering techniques were applied to the second and third harmonics generated by the 21 gesturers of the 'Take Mug' gestures, as visual interpretation of the data in Table 7.3 suggests that these harmonics control the sub-group characteristics. A careful check of the distance metric and linkage method showed some surprising differences when the vector data had been normalised by the criteria of determining the largest component of the positive or negative sequence component. In chapter 6, Table 6.6 showed clearly the advantage of using the 'ward' linkage method as the 'cophenetic correlation coefficient' was higher than using any other linking method. When the same experiment was conducted with data from the revised first harmonic normalisation technique there was a major change in the coefficient values obtained with the 'ward' linkage method. This experiment was undertaken for each of the first three harmonic components and showed the same decrease in the coefficient values as shown in Tables 7.4, 7.5 and 7.6, respectively.

Linkage/ Distance	**'single'**	**'complete'**	**'average'**	**'centroid'**	**'ward'**
City Block	0.671	0.577	0.678	**0.683**	0.3524
Euclidean	0.622	0.602	0.659	0.660	0.364
Mahalanobis	0.544	0.379	0.555	N/A	N/A

Table 7.4 Comparison of distance metrics and linkage methods by the cophenetic correlation coefficient with the revised normalisation method for the first harmonic. The largest coefficient is shown in bold.

Linkage/ Distance	**'single'**	**'complete'**	**'average'**	**'centroid'**	**'ward'**
City Block	0.574	0.524	0.511	0.586	0.461
Euclidean	0.629	0.475	0.584	0.583	0.477
Mahalanobis	0.627	0.649	**0.706**	N/A	N/A

Table 7.5 Comparison of distance metrics and linkage methods by the cophenetic correlation coefficient with the revised normalisation method for the second harmonic. The largest coefficient is shown in bold.

Linkage/ Distance	'single'	'complete'	'average'	'centroid'	'ward'
City Block	0.535	0.466	0.539	0.546	0.415
Euclidean	0.558	0.465	**0.563**	0.561	0.418
Mahalanobis	0.549	0.480	0.557	N/A	N/A

Table 7.6 Comparison of distance metrics and linkage methods by the cophenetic correlation coefficient with the revised normalisation method for the third harmonic. The largest coefficient is shown in bold.

Figure 7.11 shows the dendrogram obtained from second harmonic data using the 'Euclid' metric and 'single' linkage method. It shows no clear distinction between the links and so clusters are hard to visualise. Figure 7.12 is a graph showing how the vectors are clustered into 4 groups. The majority are grouped into cluster 1 (red), two are in the second cluster (green) and there are two single points that are allocated as cluster 3 and cluster 4, respectively. However, studying the dendrograms for links of inconsistency and consistency show the clustering is more clearly revealed with the 'ward' linkage method.

When the clustering uses the 'Euclid' metric and the 'ward' linkage method the data is divided into more distinct groups. The dendrogram and the distribution of clusters are shown in Figure 7.13 and Figure 7.14 respectively. The latter figure shows four clusters, the first and third cluster constitute the largest clusters and clusters 2 and 4 represent the pairs of outlier vectors.

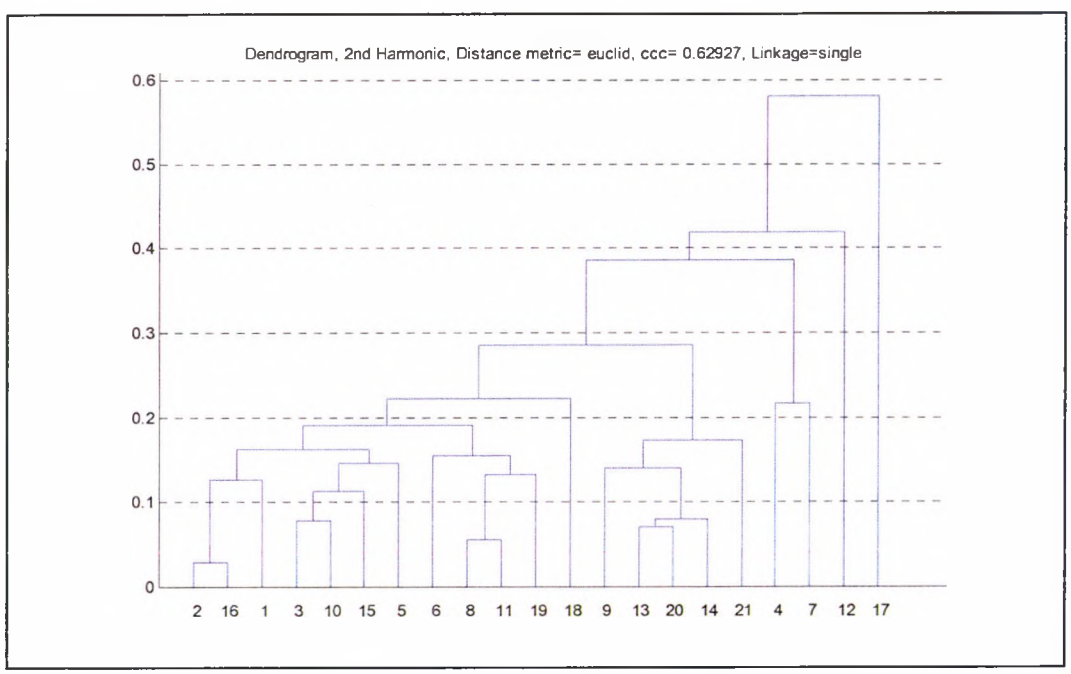


Figure 7.11 Dendrogram of the second harmonic using the Euclid metric and single linkage.

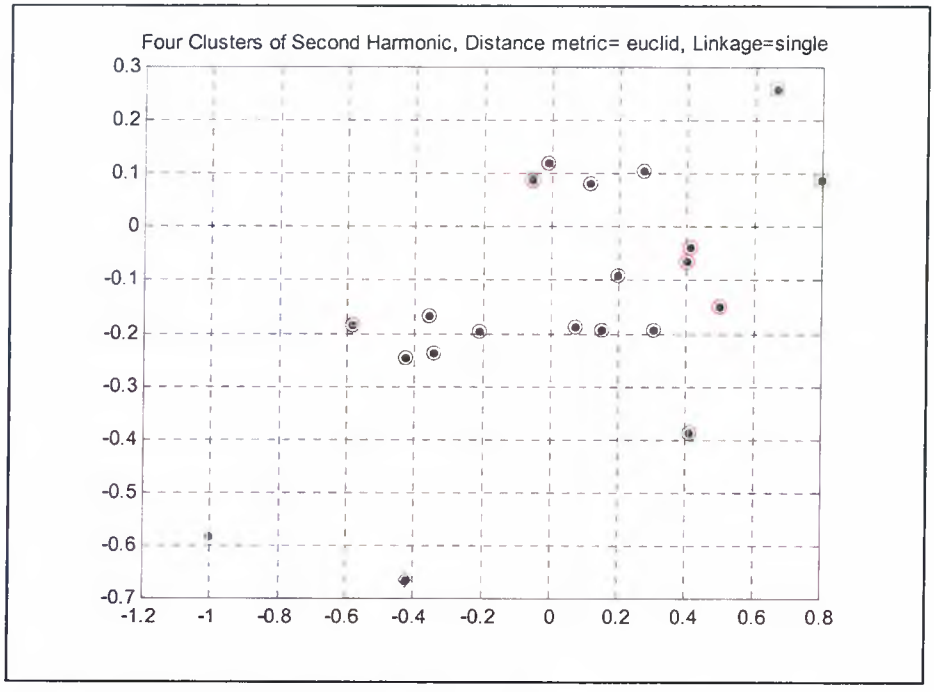


Figure 7.12 Graph showing the classification of vectors (red/circle, blue/diamond, green/square and cyan/star) using the Euclid metric and single linkage method for the second harmonic

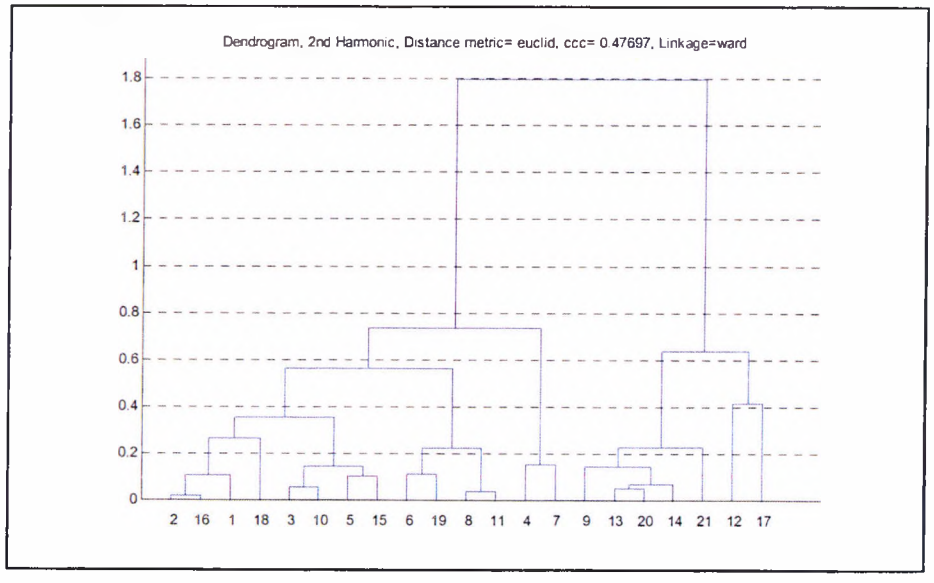


Figure7.13 Dendrogram of clusters for the second harmonic using the 'euclid' metric and 'ward' linkage method.

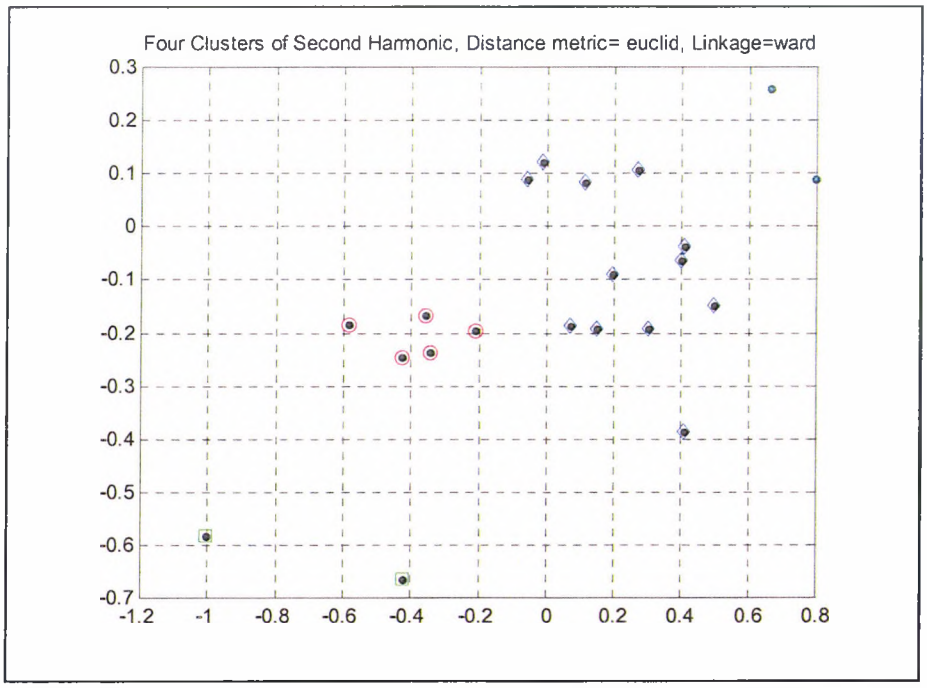


Figure7.14 Graph showing the classification of vectors (red/circle, blue/diamond, green/square and cyan/star) using the Euclid metric and ward linkage method for the second harmonic

It is interesting to note that the City Block and Euclidean metrics give virtually identical results for the various linkage methods with some unusual linking of some of the single outlier vectors (Appendix VIII).

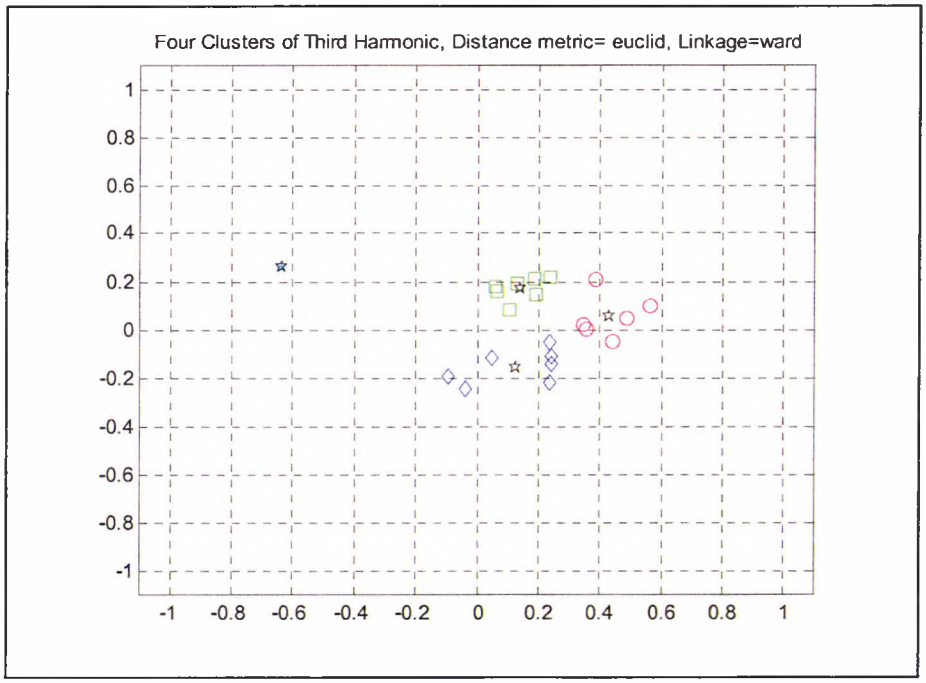


Figure 7.15 Graph showing the classification of vectors (red/circle, blue/diamond, green/square and cyan/star) for the Euclid distance metric and ward linkage for the third harmonic and showing average values of each cluster (black/star).

The distribution of clusters for the third harmonic resulted in groupings using the Euclid distance metric and ward linkage is shown in Figure 7.15. In this diagram the average of the coordinate position is shown by the star symbol.

In the first analysis of the 'Take Mug' sequences four characteristics were manually identified from the suite of twenty-one gesturers. These differences seemed to be identified by mainly the second and third harmonics. An initial approach to determining how the harmonics related to the number of characteristic in the suite of gestures was to assume that there were two distinct groupings in the each of the second and third harmonics which results in four permutations of second and third harmonic clusters. However, the results of clustering single harmonics indicated that the data did not neatly fall into two clusters because of the preponderance of outlier type of data, so there were typically four clusters per harmonic. Consequently, analysis was undertaken using four clusters from each of the two harmonics, indicating that there are sixteen possible Target Classes. The second and third harmonic cluster numbers are decoded to become one of 16 possible numbers. For example in table 7.7, gesture B has assigned to it cluster 3, of the second harmonic, and cluster 1 of the third harmonic which decodes to Target Class 9.

Gesturer	Gesturer No.	2^{nd} Harmonic Cluster Number	3^{rd} Harmonic Cluster Number	TC (Target Classes)	T (Target)
A	1	3	1	9	1
B	2	3	1	9	1
C	3	3	2	10	2
D	4	4	3	13	5
E	5	3	2	10	2
F	6	3	3	11	3
G	7	4	3	13	5
H	8	3	3	11	3
J	9	1	3	3	US
K	10	3	2	10	2
L	11	3	2	10	2
M	12	2	1	5	US
N	13	1	2	2	4
O	14	1	2	2	4
P	15	3	3	11	3
Q	16	3	1	9	1
R	17	2	4	8	US
S	18	3	1	9	1
T	19	3	3	11	3
U	20	1	1	1	US
V	21	1	2	2	4

Table 7.7 Target Classes generated for the second and third harmonics clusters using the 'euclid' distance metric and 'ward' linkage method, resulting in Five Target Classes (US=Unspecified as is a single entity)

Table 7.7 represents the clustering due to the 'Euclid' metric and 'ward' linkage method. The target classes are then analysed and the number of occurrences of each Target Class recorded. Analysis of the target classes showed only 9 of the possible 16 Target Classes used. The occurrence of these Target Classes were found to have grouped together as 5 distinct classes as shown in Table 7.8. The five groupings of Target Classes are reassigned so that for example the four gesturers 1,2,16 and 18 are classified as having similar second and third harmonic properties which were decoded as Target Class 9 and now assigned as Target Number 1 for convenience. Gestures that were unique, i.e. showed no similarity of harmonic content with other gestures, were ignored in the analysis and shown as US (Unspecified in the table).

Target Number	Target Class Number	Gesturer Number	Second harmonic Cluster No.	Third Harmonic Cluster No.
1	9	1,2,16,18	3	1
2	10	3,5,10,11	3	2
3	11	6,8,15,19	3	3
4	2	13,14,21	1	2
5	13	4,7	4	3

Table 7.8 Target and Target Classes and associated cluster of the Second and Third harmonic using the Euclidean distance metric.

The next step is to use the Targets in a PNN network to compare classification results. In order to use the Target information for the PNN the average position for each cluster is determined for each harmonic. For example a 'star' marks the average value for each cluster for the third harmonic data in Figure 7.15. The coordinates and equivalent modulus and angle for Euclidean distance metrics, used to realise Targets, are shown for the second and third harmonic are shown in Table 7.9.

	Coordinates -real	Coordinates -imaginary	modulus	angle
Second harmonic				
Target 1	-0.3862	-0.2051	0.44	-152
Target 2	*-0.716*	*-0.624*	*0.950*	*-139*
Target 3	0.228	-0.0752	0.24	-18
Target 4	0.7296	0.1747	0.75	13
Third harmonic				
Target 1	0.4273	0.0549	0.43	7
Target 2	0.1387	0.1686	0.22	-51
Target 3	0.1238	-0.1566	0.20	-52
Target 4	*-0.6378*	*0.259*	*0.689*	*158*

Table 7.9 Target coordinate data – Euclidean distance metric (Italic data not used)

Comparison of the different classification techniques are shown in Table 7.10. Classification of gestures using the PNN (with targets of gestures A, G, K and M) gave very similar results in nineteen of the twenty-one gestures, as to visual methods. Gestures L(11) and P(15) were classified differently, but the next nearest target in

either case would have made the classification for the two columns the same. This was somewhat surprising as there were distinct differences in the clustering of the second harmonic data for the two distance metrics.

Gesturer	Gesturer Number	Visual classification	Classification using A, G, K, M as Targets	Target Classification (Euclid metric)
A	1	α	1	1
B	2	α	1	1
C	3	γ	3	2
D	4	β	2	3
E	5	γ	3	2
F	6	β	2	4
G	7	β	2	3
H	8	γ	3	4
J	9	γ	3	5
K	10	γ	3	2
L	11	γ	3	**2/4**
M	12	δ	4	5
N	13	γ	3	5
O	14	γ	3	5
P	15	γ	3	**4/3**
Q	16	γ	3	1
R	17	δ	4	5
S	18	α	1	1
T	19	γ	3	4
U	20	δ	4	5
V	21	δ	4	5

Table 7.10 Classification of twenty-one 'Take Mug' gestures using the PNN/ clustering technique) using Euclidean distance metric and 'ward' linkage method and compared to the visual classification (bold shows the difference in the metric methods and / show the next nearest target).

There is also much similarity between the visually clustered grouping and those obtained using the clustering method. The main difference being that the visual technique gave 4 sub-groupings, whereas the clustering technique gave 5 sub-groupings. This is more clearly shown in Table 7.11, where the individual gestures are grouped together according to the classification technique.

Characteristic 'α' was virtually the same by the cluster method and the visual inspection method. The hand took the imaginary mug and then went directly down to the resting place, in a manner similar to the avatar example, without returning the mug from the imaginary table. Gesture Q (16) is now included in this sub-grouping, and was perhaps overlooked in the original visual classification.

Characteristic 'β' was characterised by the deliberate way the imaginary mug was picked up, and gestures D(4) and G(7) are common to all three classification methods. It is worth noting that the characteristic 'δ' that was described as a 'wide'

characteristic. It is common to all three classification methods, but the PNN/clustering technique has added three more gestures (9, 13, 14) to the sub-group, which visually had previously been allocated to the 'γ' characteristic. This final grouping has been split into two distinct grouping by the clustering technique.

Characteristic	PNN A, G, K, M targets	PNN Clustering derived targets
α	1, 2, 18	1, 2, 18, **16**
β	4, 7, **6**	4, 7
δ	12, 17, 20, 21	12, 17, 20, 21, **9, 13, 14**
γ1	3 ,5 ,10 ,11, **9, 13, 14**, , 16, **8, 15, 19**	3 ,5 ,10 ,11
γ2		6, **8, 15, 19**

Table 7.11 Comparison of classification techniques for 21 'Take Mug' gestures (bold shows difference between the two classifications techniques)

In general the PNN/clustering method has shown the same divisions as the visual method. The main difference has been that the characteristic 'γ' has been further sub-divided with some gestures being classified as being closer to characteristic 'δ' in nature and then splitting the remaining gestures into two groups, 'γ1' and 'γ2'. It is highly likely that the PNN/clustering technique has recognised characteristics of the 'Take Mug' gesture better than recognition made by the human observation. The grouping of characteristic 'γ' was rather large (over half the gestures) indicating that it was difficult to visually distinguish between characteristics.

7.4. Gesture Stimuli Experiments

The 'Take-Mug' gesture was an arbitrarily chosen gesture action that showed how the properties of the first three harmonics could be used to characterise and distinguish between several people undertaking the same gesture. The Gesture Stimuli experiments were selected in consultation with Woll (2004). These gestures are often used in 'normal' communication and also by disabled people to help communication, but have often been disregarded because of the difficulty of describing them by objective means. As explained in Chapter One, relevant non-verbal communication, such as gestures, associated with verbal communication can have a greater impact on the recipient than just verbal communications. There is also evidence that this type of gesturing was a precursor to spoken language. But the resulting gestures are more complicated in trajectory path and frequency content than simple pointing gestures.

7.4.1. The Gesture Stimuli Experiments

The gesture stimuli experiment involved recording fifteen people, mostly students aged under twenty-five, but there were representatives from their forties, fifties, seventies and eighties age group. About two-thirds of the people were Caucasian and the remainder were of Asian decent. Consultation with Woll (2004) suggested a range of stimuli that had been used with disabled people. Nine different gesture

stimuli were recorded for each person. Picture outlines of objects were prepared on a set of stimulus cards. The cards had outline drawings of a toothbrush, a knife, key, screwdriver, hairbrush, hammer, whisk, hand-saw and bottle-opener. An example of what was expected for one stimulus was given before the experiment commenced. Each person was asked to interpret into a gesture action what he or she was shown on the card's 'picture'. The question, 'What would you do with?' was asked as they were shown the card before they performed the action associated with the stimuli. It was expected that all gestures would be single-handed and predominantly right-handed.

Two different environments were used for the recording, an office (Figure 7.19) and a room in a home (Figure 7.22). From a recording point of view the lighting conditions in each were not optimal. In the office situation sunlight was restricted from entering the office by Venetian blinds, but its strength propagated through to make magnolia coloured walls look pink, although some of this effect may have been caused by the automatic white balance compensation mechanism of the camera. This resulted in the skin-coloured regions being difficult to segment from the wall region. The Hue range being recorded at –0.19 to 0.095 (or -0.05), indicates that white balance is not correct as the Hue has moved into the magenta range. The saturation range improved the segmentation and was restricted to the 0.1 to 0.34 range.

Figure 7.16 A challenging environment for skin-coloured segmentation with inappropriate SCM objects recorded by the green and blue crosses.

The image sequence was affected by noise and as previously discussed in Chapter 4, the red, green and blue crosses seen in Figure 7.16 and 7.19 show the positions of the rank-ordered three most significant skin-coloured moving objects. The most significant object (red) correctly relates to the position of the hand. The green and blue crosses relate to stationary objects. These occurrences were more prevalent in some sequences than others and may have been caused by light changes or compensation changes taking place whilst recording, or due to the recording process. Experience with the home environment, although not well lit, did not provide so many spurious gesture objects and so tracking performed better.

Figure 7.17 shows a typical image sequence and a relatively large number of second most significant gesture objects (green 'o') because of the noisy conditions. The output of the 'OSA' is shown by the black crosses. This output is then prepared for frequency analysis by finding the starting and stopping places of the gesture, as shown in Figure 7.18. Because of the poor lighting conditions and sporadic false

objects some sequences had to be edited to improve the output ready for frequency analysis. The position when the tracking algorithm took the wrong course was noted and was steered into the correct path. This occurred typically about once every other gesture.

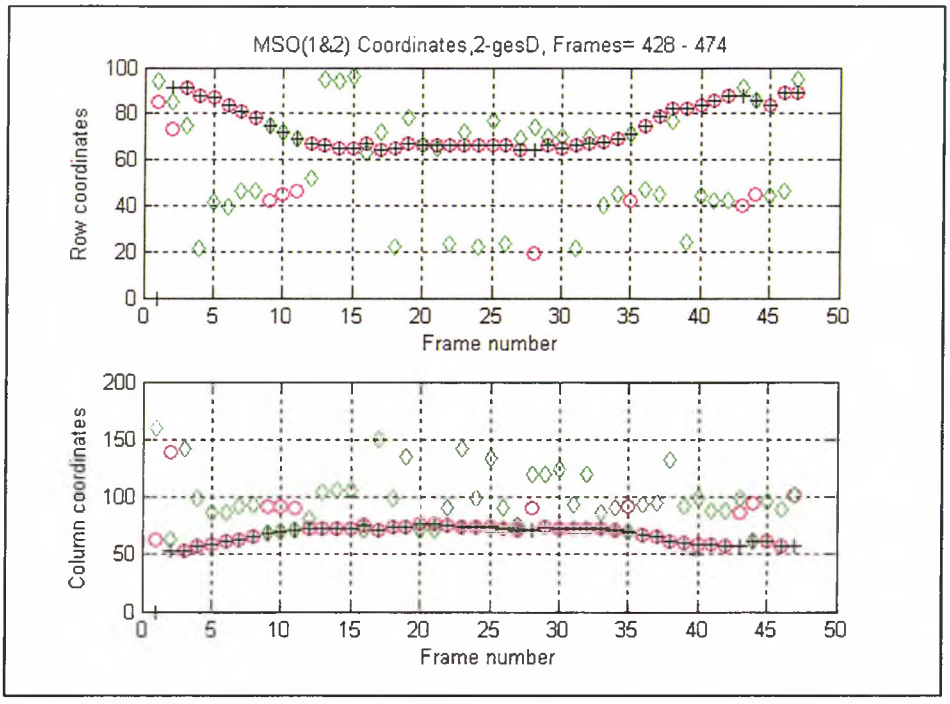


Figure 7.17 The OSA output, '+', chosen from the two most significant SCM objects (red and green respectively)

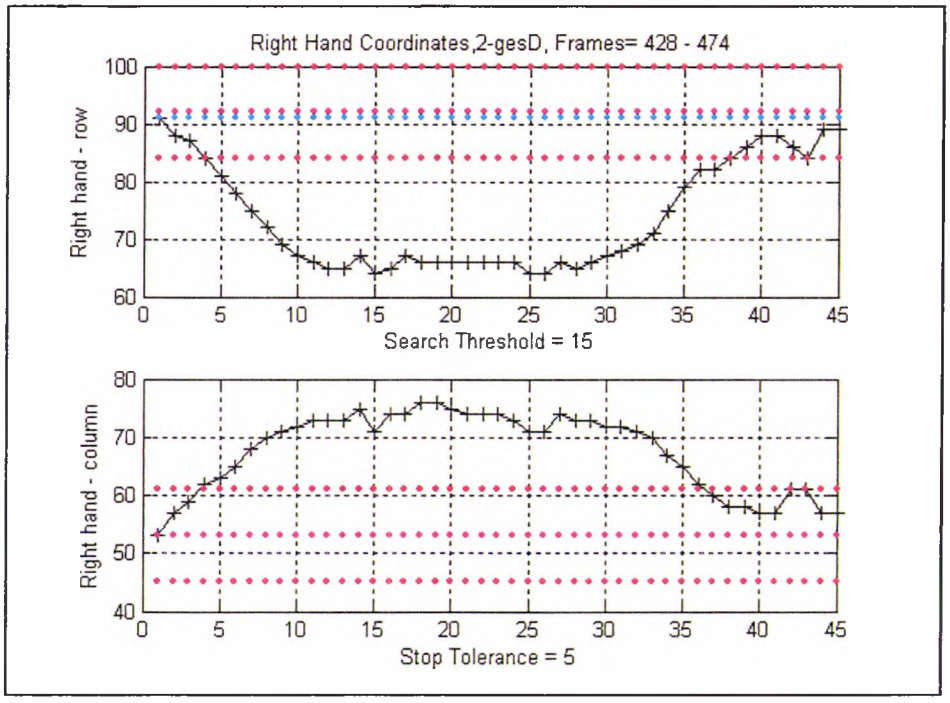


Figure 7.18 Segmented gesture ready for frequency analysis, showing right hand row and column coordinates with initial search region (red dots), start coordinates (cyan dots) and stop condition (magenta dots).

In preparation of the data for frequency analysis the search threshold also had to be adjusted. In most cases when the hand is gesturing the velocity of the hand does not change unduly. Tracking can take place with a simple threshold and can follow the most significant object with little error. With the poorer lighting conditions, the number of possible other objects to follow increases and wrong tracks can be more easily followed. One situation where errors can easily be made is when the hand goes in or out of the 'strike' phase. The velocity of the hand rising or falling can be quite extreme. The standard frame rate is almost insufficient to capture the hand's trajectory with sufficient samples. The outcome is that the distance moved in this mode can exceed the threshold, so no suitable tracking object is found or an erroneous object is selected and an incorrect track is followed.

Poor segmentation also caused problems in distinguishing the hand position when close to the face or close to the other hand. Some problems were noted with the hand not returning to the correct position or falsely triggering the stop condition by passing through the stopping coordinate window, although the gesture had not finished. It was also noted that the gesturer was anticipating the initial hand movement of some gestures as they became familiar with the experiment. This meant that the dominant hand did not return to the starting point (typically on the leg), but was held higher at chest level in anticipation of the next action. The coordinates of this position would then be outside the initial search area and cause a truncation problem at the beginning of the gesture rather than at the end of the gesture.

The more simple gestures as in the previous example (Figure 7.17) were 45 samples long, others as in Figure 7.19 and Figure 7.22 were 83 to 112 samples long, respectively. For those samples exceeding 60 samples in length frequency analysis was conducted on alternate samples. Many of the actions were similar from gesturer to gesturer, but there were many idiosyncrasies in the interpretation too. For instance, some of the teeth cleaning gesturing were undertaken as if in the real situation and the gesture length approached 180 samples.

The experiment was quite ambitious in its size and scope and some of the gesture sequences were not considered reliable, without ground truth data to verify the sequences. There were a number of reasons for this, including tracking errors, although some sequences and results could be validated. These are described in the next section. The new limitations, to the methods used before with the 'Take-Mug' sequences, are also explained here.

7.4.2. Frequency Analysis of Gesture Stimuli

Figure 7.19 shows the environmental conditions with a gesturer performing the 'whisk' action. An interesting consequence of the analysis of the gesture stimuli is oscillatory components captured in the gesture, as can be seen in the tracking data of Figure 7.20.

The automatic method of generating gesture objects for hand tracking was used. Ground truth data was not generated for these experiments but the action in the image sequence could be seen to coincide with the captured row or column oscillations. The complete sequence of images (83 frames) is shown in Appendix VII and relates to the row and column coordinates in Figure 7.20. In this figure the high

oscillatory action is apparent in the column data, but does not appear in the row data as the data is relatively constant.

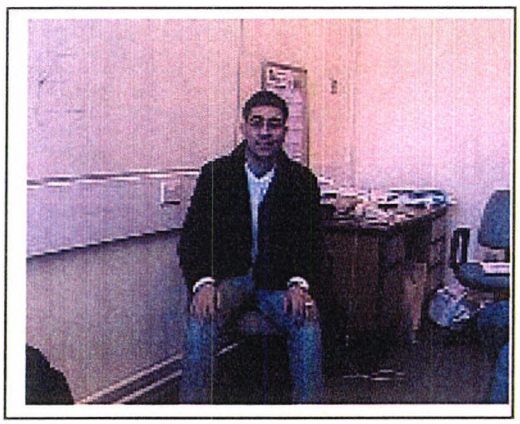


Figure 7.19 Image from the 'whisk' sequence showing gesturer and environmental conditions

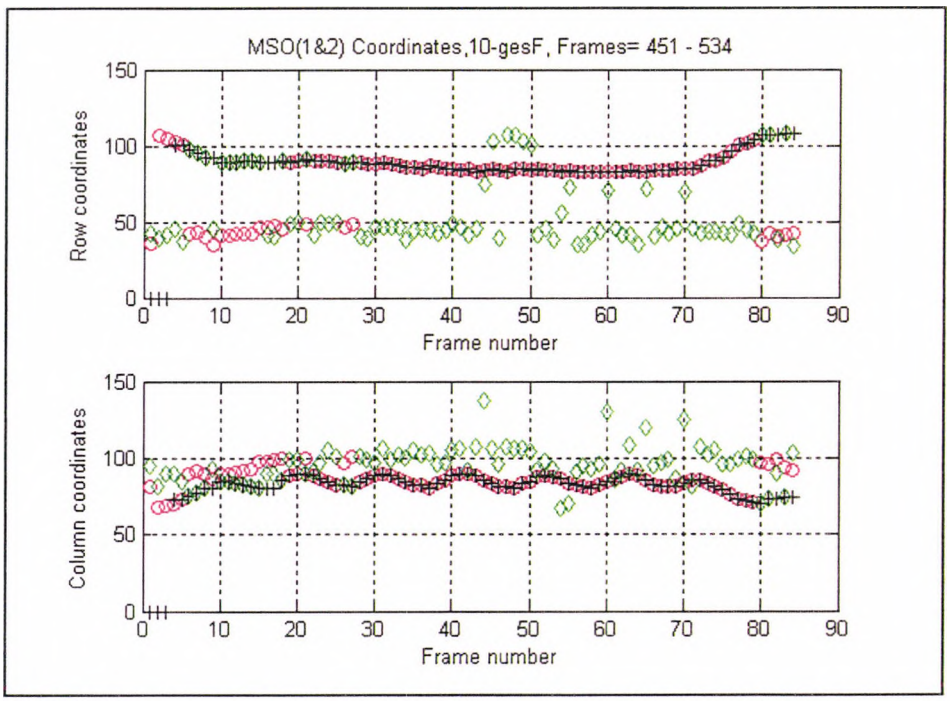


Figure 7.20 The trajectory coordinates due to the 'whisk' action

These observations are shown much clearer in the change of scale in figure 7.21 that show the data prepared for frequency analysis. There are seven distinct oscillations in the gesture period. This observation is confirmed by the frequency response results shown in the Table 7.12, where the amplitude of the positive and negative sequence of the seventh harmonic are both very similar at about 0.68 compared to the normalised amplitude of unity for the first harmonic positive sequence component. This amplitude is much larger than usual as the amplitudes usually decrease as harmonics increase

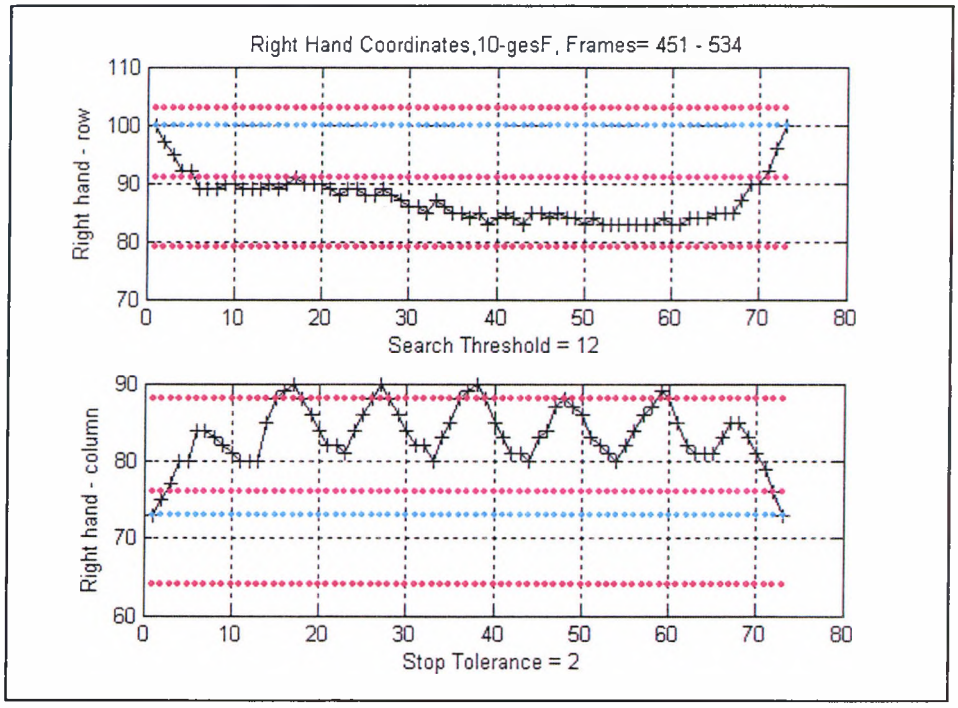


Figure 7.21 'Whisk' gesture coordinates prepared for frequency analysis

Harmonic	Positive Magnitude	Positive ϕ	Negative Magnitude	Negative ϕ	Orientation θ
1	1.000	39	0.449	-39	20
2	0.363	11	0.343	-11	35
3	0.283	-2	0.444	2	-18
4	0.149	-17	0.151	17	-19
5	0.166	3	0.169	-3	-18
6	0.178	17	0.073	-17	-65
7	**0.678**	-5	**0.689**	5	152
8	0.015	59	0.092	-59	7
9	0.099	12	0.063	-12	-119
10	0.009	22	0.112	-22	-50
11	0.039	0	0.062	0	6
12	0.079	119	0.085	-119	66

Table 7.12 'Whisk' frequency content, showing the prominence of the seventh harmonic (bold).

Another oscillatory result is shown in Figures 7.22 and shows different environmental conditions and gesturer than before used to capture a 'saw-action' gesture. The complete sequence of images (112 frames) is shown in Appendix VII. Figure 7.23 and Table 7.13 show the four very deliberate 'saw-action' harmonics. It is interesting to see that in the 2D image of Figure 7.24 the relatively large amplitude

of the 4th component most strikingly at an orientation angle different to the other first three harmonics, which are almost in line with each other.

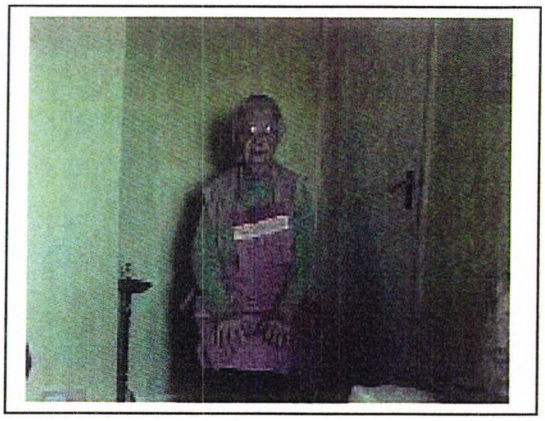


Figure 7.22 Image from the 'saw-action' sequence showing gesturer and environmental conditions

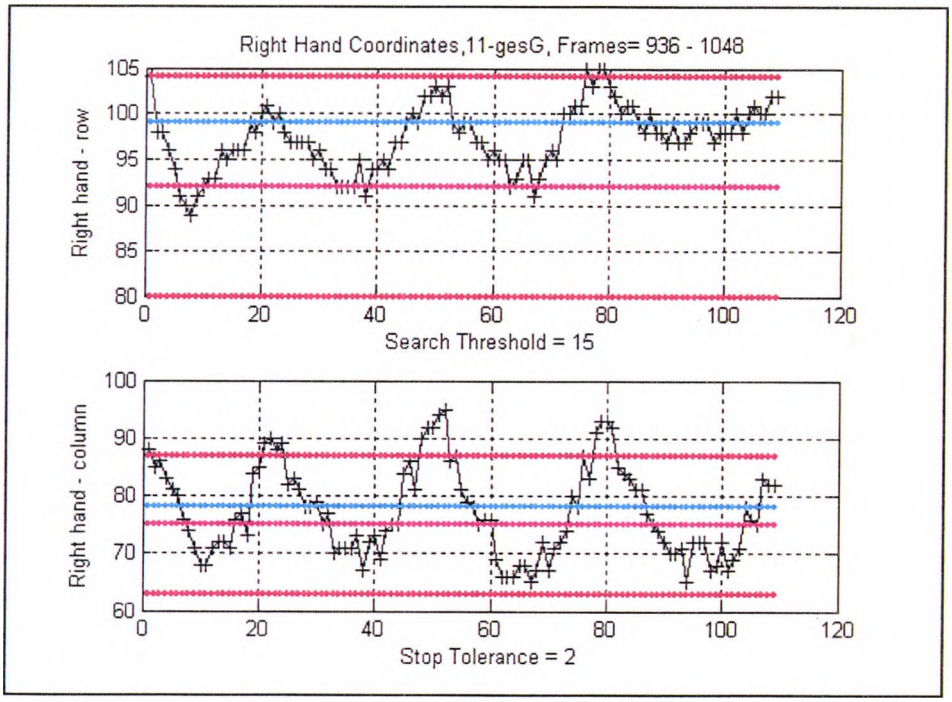


Figure 7.23 'Saw-action' gesture coordinates prepared for frequency analysis

There were several observations to be made from the results of the frequency analysis of the gesture stimuli experiments. Firstly, it was noted that some gestures were much longer than previously observed and enacted most vigorously by the gesturer. One gesture was recorded at making in excess of twelve oscillations in one gesture. This would require the default setting of twelve harmonics to be reconsidered for such activities.

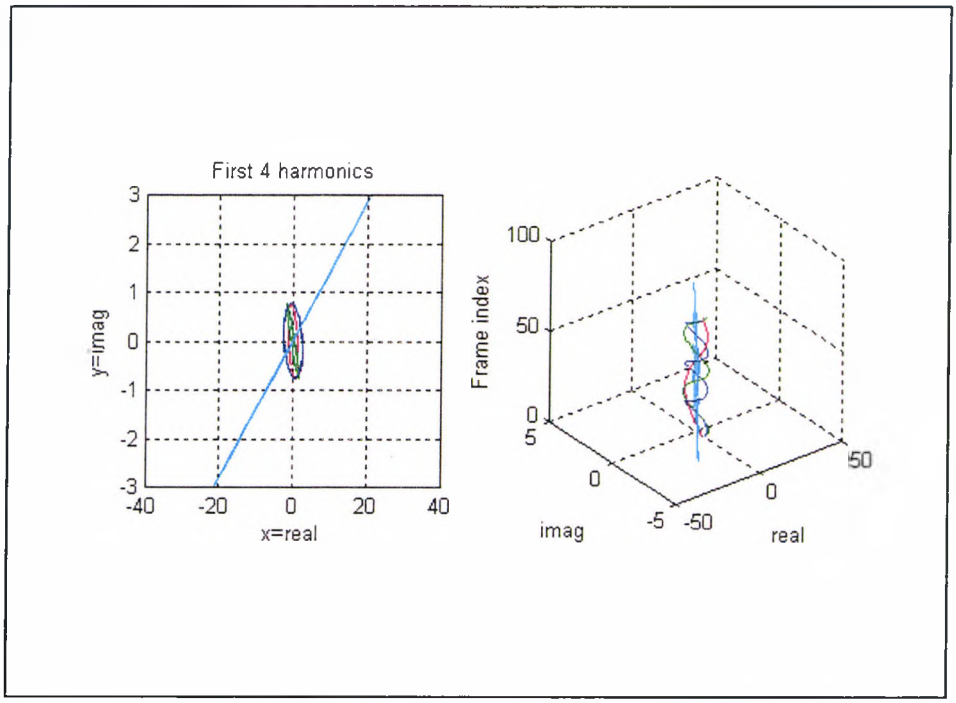


Figure 7.24 2D and 2DT views of the frequency components of the 'Saw-action'. The 'cyan' 4th harmonic component is shown as much greater than the other components.

.Harmonic	Positive Magnitude	Positive ϕ	Negative Magnitude	Negative ϕ	Orientation θ
1	0.4099	-105	1.000	105	16
2	0.8017	16	0.983	-16	6
3	1.5261	86	0.841	-86	-18
4	**9.7202**	-117	**9.722**	117	-12
5	0.4626	-50	0.886	50	-7
6	0.7929	-26	0.966	26	-4
7	0.5316	166	0.507	-166	21
8	2.9921	-30	2.852	30	-9
9	0.2736	18	0.456	-18	112
10	0.4851	-13	0.904	13	-65
11	0.9979	-20	0.896	20	-46
12	0.7474	168	0.626	-168	2

Table 7.13 'Saw-action' frequency content, showing the prominence of the fourth harmonic (bold)

It is evident that for many gestures the normal association of harmonic amplitude decreasing as frequency value increased held good. However, with the advent of an oscillatory activity the oscillatory amplitude component could be greater than the normally observed value as in the two examples of the 'whisk' and the 'saw-action'. Without the whisk action the 7th harmonic amplitude would have been expected to have amplitude of about 0.16 instead of the 0.68 recorded. Whereas with the 'saw-

action' action the fourth harmonic is some nine times greater than the amplitude of the first harmonic. Strategies could be put into place to automatically recover this dominant oscillation. A possible solution would be to use the clustering technique as the vector for the oscillatory frequency would appear as an outlier in the analysis. However, there would be no guarantee that a gesturer would repeat the gesture again with the same number of oscillations. This could therefore be a fruitful area of further research.

By contrast the performance of the tracking system became less robust as the amplitude of the gesture became less. With the lower amplitude oscillation the amplitude became similar to the uncertainty in hand position due to variability in gesture object size as discussed in Chapter 3. In addition, the gestures with smaller oscillatory actions generally made the tracking less certain.

More oscillatory components in the gesture trajectory stopped the gesture from being modelled by just three harmonics as in the 'Take-Mug' gesture experiments. An example shown in the Figure 7.25 shows the original row and column trajectory data (cyan 'o') with black '+' showing the reconstruction of the waveform from frequency analysis with six harmonics. When the reconstruction is reduced to four harmonics (red '+') noticeable differences are apparent from the original trajectory. It is interesting to note the highly elliptical nature of each of the harmonic components (Figure 7.26) which is confirmed by the tabular results (Table 7.14) that the high frequency amplitudes are greater than would be expected.

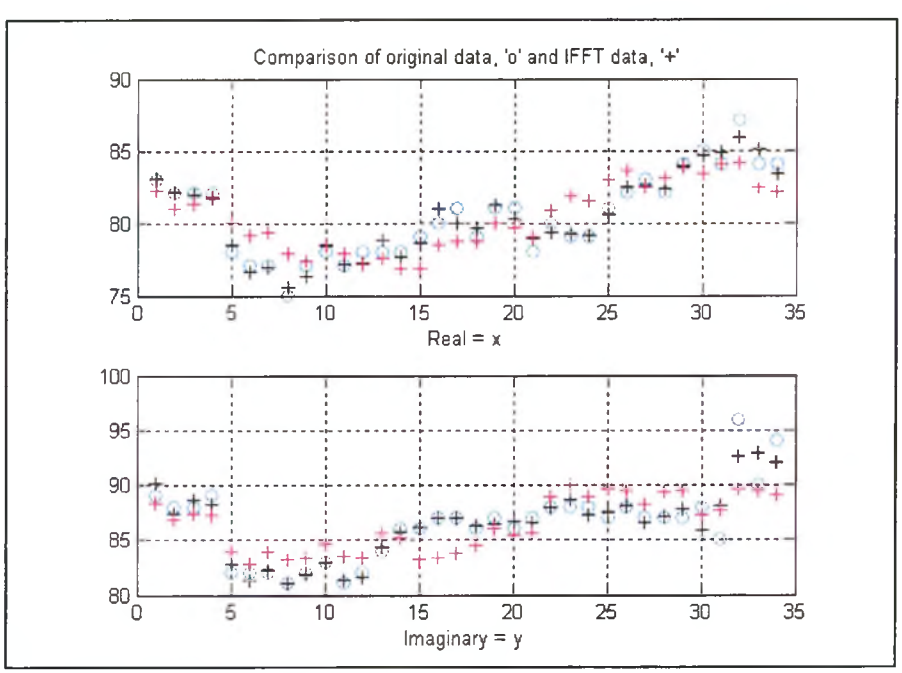


Figure 7.25 A gestures coordinates (cyan, 'o') with IFFT reconstruction using 6 (black '+') and 4 (red '+') harmonics

It is also noticeable that in this particular example that the harmonic amplitudes from the fourth harmonic to the twelve harmonic are significant. In other examples the higher order harmonics have decreased in amplitude with frequency and have allowed gestures to be described with just the first three harmonics. The exception was the case when there is a definite action as in the 'whisk' and 'saw-action'

gestures when significant higher harmonic amplitude is captured. It is concluded that as oscillatory amplitude becomes less in a gesture and gesture object realisation becomes less reliable (objects generated at various parts of the hand, so introduces competing higher oscillatory components into the data) the tracking and analysis becomes tenuous.

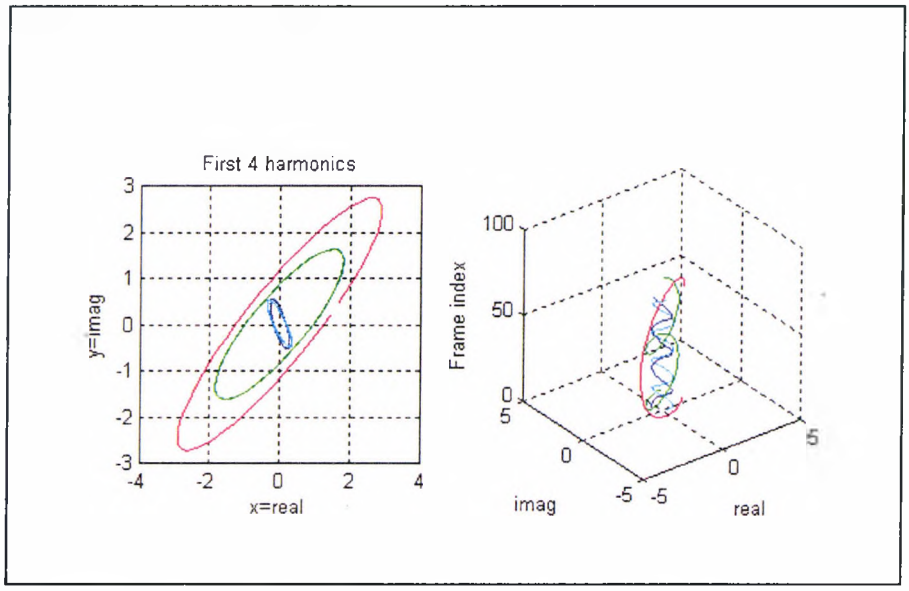


Figure 7.26 2D and 2DT views of the first four frequency components of a gesture with low amplitude oscillation.

Harmonic	Positive Magnitude	Positive ϕ	Negative Magnitude	Negative ϕ	Orientation θ
1	1.00	-72.5	1.58	72.5	-42.99
2	1.00	-38.9	0.55	38.9	1.99
3	0.13	22.2	0.25	-22.2	-72.8
4	0.25	50.9	0.14	-50.9	169.9
5	0.23	-34.4	0.35	34.4	79.2
6	0.46	-19.3	0.44	19.3	-15.1
7	0.46	-18	0.34	18	-17.6
8	0.30	-40,6	0.31	40.6	7.7
9	0.32	-46.6	0.36	46.6	-9.8
10	0.26	-48.3	0.38	48.3	26.7
11	0.29	42.0	0.15	-42.0	-194
12	0.29	39.0	0.28	-39.9	-9.35

Table 7.14 Frequency content of a low amplitude oscillation gesture.

The interpretation of these gestures tended to be more freely interpreted and so the first harmonic orientation angle took on many different values for the same gesture by different gesturers. Although many sequences of gesture were recorded it became apparent that each gesture was unique because of the lack of any physical guide or constraints as with the 'PETS', 'Take Mug' or 'Pointing' experiments. The results of the analysis of the gesture stimuli showed that each gesturer's response was so varied

that any comparison between gesturers was not worthwhile. In fact a series of results of individual's actions with gesture stimuli would probably give a unique signature of the individual.

7.4.3. Additional Observations

It is worth recording a number of additional observations that were made with the gesture stimuli experiments. These observations are included as they are perhaps worthy of note to pursue in an alternative endeavour. They are not followed through scientifically, but may need to be considered as necessary when reviewing the nature, behaviour and interpretation of gesture activity.

Visual inspection of the image sequences resulted in some interesting and unexpected findings. The right hand was the dominant hand for thirteen of the fifteen people. One person was left-handed and one was ambidextrous. Interestingly, the ambidextrous gesturer found it difficult to do the task without explaining what was being undertaken.

Most gestures used single arm movement of the dominant hand. However, some showed the less-dominant hand moving in synchronism with the dominant hand for a short while. For instance most people, when asked to saw, held an imaginary piece of wood with the less-dominant hand, whilst showing sawing action with the dominant hand. Similarly the bottle opener action often showed the less-dominant hand being used to hold the imaginary bottle.

Observations of the responses showed that these mimetic gestures lacked the precision of other types of gesture that are well defined. Pointing to the right or the left is well understood by the majority of people so the variations between people are kept to a minimum. The individual gesture stimuli show some similarities between people but overall each gesture has a distinct 'signature'. It had been expected that the gestures would have been quite short. For instance, when showing hair brushing just a couple of generic strokes would have been recorded. However, what was recorded was the duration and the much personalised ritual that the gesturer undertook as if the tasks were actual.

The misinterpretation of two of the tasks, by two of the older people, brings into question the role of personal background and experience of the gesturer. This affects their anticipation of what is expected in the gesture by the observer. Misinterpretation could also be due to a poor drawing, and maybe photographic pictures should be used instead to reduce this possibility of error. It was also noticed that some gesture stimuli left the gesturer quite perplexed and unsure as to how to act. For example, the bottle opener required many to hesitate and think of the use and context of the action before deciding on whether to use one or two hands.

This series of experiments showed that 'normal' people have problems interpreting gesture stimuli. The actions appear to rely on past experience as to the interpretation of how to use the object. Maybe more significantly the stylised response was completely unexpected, which may have a variety of psychological explanations that made the output so variable. The key conclusion is that the mimetic gesture lacked precise definition, so comparisons are likely to be difficult. These actions may be a

unique extension of the gesturer so each gesture is a unique quality of every individual.

7.5. Summary

This chapter shows how Fourier analysis, combined with PNN and cluster techniques, is a powerful technique for gesture recognition. Both the 'PETS' hand raising gesture and the 'Take Mug' gesture show that the first harmonic orientation angle to be closely clustered regardless of one person repeating a gesture or many people repeating a gesture. This is because the start/end of the trajectory and the height (approximately mid-point) of the trajectory are related to relative spatial dimensions. These two groups of experiments also demonstrate how intra-class and inter-class differences can be recognised.

The gesture stimuli experiments showed that the interpretation of the gesture stimuli were not constrained by human body attributes and so the first orientation angle was not tightly clustered. However, the harmonics do capture the unique characteristic of a particular gesture by a gesturer and could be used in a recognition system that has a large database of possible gestures and gesturers.

In the case of a gesture that has the first harmonic tightly clustered, the other harmonics can vary widely. It was shown that just the second and third harmonics could describe a range of differences in a repeating gesture. An experiment was conducted with the 'Take Mug' gesture that showed that PNNs with target gestures obtained from clustering technique could distinguish a range of differences in the same gesture as well as, and if not better than, visual classification. This resulted in five main groupings being classified whereas four had been classified visually.

Further revelations were made about the distance metrics and linkage methods that are used in the clustering techniques. In chapter six it had become clear that deciding on what distance metric or linkage method to use was not clear cut. However, when the frequency data was normalised on the basis of the largest of the positive and negative frequency components the cophenetic correlation coefficient did not give a reasonable indicator of which to use. However, study of the dendrograms and the inconsistent coefficient still indicated for this type of data that the 'ward' linkage appeared to group the data in the most sensible manner. The City Block and Euclidean distance metrics were quite similar in their action. It was interesting to note that the 'single' or 'nearest neighbour' linkage technique located the outlier gesture better than the 'ward' technique, when a representative first harmonic vector was required, albeit that the vectors were closely clustered within experimental error.

The duration of the gesture stimuli gestures was much longer than expected so sub-sampling by a factor of two was introduced for samples greater than sixty. When gestures were relatively uncomplicated in their trajectory sub-sampling was used with out any detrimental effect on results. Results showed that there was virtually no difference in harmonic components as a result of this action. The most interesting observation with the gesture stimuli experiments was the number of gesturers that showed repeated hand movement and hence oscillations in the data. This data was quite conspicuous, as the amplitude of the harmonic did not fit the normal decreasing amplitude with frequency profile. Oscillations also tended to be whole cycles due to

natural tendency for arms to rise and then fall due to gravity and there was no spectral leakage. It was noted that some gesturers had a relatively high number of oscillations in their gesture. In previous experiments an arbitrarily twelve harmonics had been used for frequency analysis as it was found that only the first few lower order harmonics were required to characterise a gesture.

In these experiments where higher oscillatory components were possible it is advisable to increase the number of harmonics above the present default level. The observation of low amplitude gesture oscillatory activity and some uncertainty as to the gesture objects position on the hand region introduced high frequency components into the harmonic components. It also made the tracking more problematic and so limited the reliability of the tracking data and subsequent frequency analysis, except when oscillatory activity was at significant amplitude.

The observation of the oscillatory nature of gesture stimuli suggests that the previous definition of gesturing needs to be modified. The action of 'beats' is said to be bi-phasic and a normal gesture to be tri-phasic. It would seem that the third component has two sub-divisions that of the intentional 'strike' part of the gesture and then the added beats or oscillatory motion. This characteristic is hardly observed in sign language gesturing. Gibet (2001) stated in Chapter 5 that apart from line, arc, static, and circle primitives only 2.8% of primitive actions can be counted as complex and described as zigzag, waves and spirals. The gesture stimuli response has shown a predisposition of gesture stimuli to have a strong oscillatory nature.

This area of gesture stimuli appears to be a much larger area of research than had been originally expected. Gesture stimuli seem to involve various degrees of symbolic, metaphoric gesturing plus the ritual and idiosyncratic actions of the gesturer. Furthermore, in some states the gesture appears to include the non-dominant hand and head in some involuntary associative action. The possibility of gesture stimuli being a diagnostic tool for certain medical conditions or as a non-invasive gait recognition tool is viable and full of potential.

8. Conclusions and Future work

8.1. Summary

The original aim of the research had been to design and test a gesture recognition system with a number of different gestures and gesturers. Implicit in the aim was the selection of cues for tracking, establishing a tracking mechanism. From this data was developed an analysis technique that could then be used for the recognition of gestures.

Various techniques have been used for the analysis of gesture. The attraction of the use of HMMs has mainly been due to its ability to deal with the variability, uncertainty and probabilistic nature of gesture. Although HMMs can be used in continuous sign language recognition systems, they need considerable training and exhibit some lack of adaptability. Some authors have suggested that a particular recognition technique is more appropriate for certain types of gesture. Similarly, others compared recognition techniques but have ignored the temporal segmentation problem which relates to the kind of gesture being studied. Recognition techniques have, however, experimented with a number of different gesturing situations including sign language, pointing gestures, mouse movements and a number of different physical activities.

The recognition technique based on Fourier analysis arose from the assimilation of several ideas. The use of frequency components was inspired by the work of Masters (1994) who used frequency components as input to a neural network for solving a number of pattern recognition problems. The advantages of using the PNN based on a RBF network came from the work of Howell and Buxton (1997): It avoided the need for extensive training. The 'glue' that linked these ideas together came from the multi-rate methods to change the sampling rate of the gesture trajectory. The gesture samples were normalised to the same number so that the frequency component could be effectively compared. This method enabled the variability in duration of a gesture to be normalised so the gesture became time-invariant.

From the observation of people gesturing, an hypothesis was made about gesturing. The hypothesis considered that in a scene with a single gesturer there were just three areas of skin-colour in motion: the two hands and the face. With single-handed gestures the dominant hand would produce the most significant movement in the image with just occasional, less significant movement being generated by the non-dominant hand and head. The cues of skin-colour and motion were investigated and fused together to form gesture or skin-coloured and motion (SCM) objects.

The work on skin colour detection centred on the use of the HSV colour space model. It was observed that Hue was affected by the illuminant and the white balance settings. The automatic white balance compensation techniques used in the video recording devices were not reliable and skin-colour could vary widely. As a consequence, before analysing an image sequence for skin-colour, a sample of the colour of the hands and forehead were taken. It was observed that hand colours may be similar, but the forehead was often different, especially if the gesturer was flushed. Skin-colour changes often resulted in the normal reddish-orange colour shifting toward the magenta colour region. Taking averages of colours that straddled

the 0-1 or 0°-360° discontinuity resulted in false values being recorded. In order to compensate for this error the RGB to HSV conversion algorithm was modified to move the discontinuity to the cyan region so that hue was measured on a plus 0.5 to minus 0.5 or a plus 180° to minus 180° range.

Motion detection was achieved by image differencing. Adaptive background techniques were found to be unnecessary as illumination changes did not affect short gesture sequences. The fusing of the skin-colour and motion cues was achieved by taking the binary masks of the skin-colour regions and difference images, using suitable threshold values, with a logical AND operation. This operation formed another mask with objects of variable size. The sorting of the objects by area invariably ranked the object associated with the dominant hand as the most significant object, followed by objects associated with the non-dominant hand, head and noise. This method also avoided the use of heuristic methods for setting the threshold levels for motion capture. These instances when the most significant object was not linked to the dominant hand the motion were small, typically at the beginning and end of the gesture or at the height of the trajectory. In this case other skin coloured motions or noise could be ranked higher than the object associated with the dominant hand. In addition, in instances when hand movement was small or in poor lighting conditions, the object related to the dominant hand was likely to disintegrate to a number of smaller objects. The issue of lighting quality in judging the effectiveness of a gesture recognition routine is yet to be addressed. There is a move to record colour temperature with publicly available gesturing sequences, but no measure is made of the quality of lighting which in this thesis has shown to affect segmentation and tracking performance.

Tracking the dominant hand was made through an object selection algorithm (OSA), that decided which object was the mostly likely to represent a region of the dominant hand, when there was more than one object to choose from. The tracking of the hand through the stroke phase of the gesture was invariably correct due to the likelihood of the most significant object being related to the dominant hand. The tracking algorithm continued to work even in some image sequences that were not well lit and complete segmentation of the hand was not possible.

The object selection algorithm was used on two different situations not included in the original skin-colour motion hypothesis. Firstly, the algorithm was used in an image sequence with three people moving in the same scene and was found to still allow the tracking of the dominant hand of one of the gesturers. Secondly, the algorithm was modified so that there were two outputs: one for the right hand and one for the left hand. Tracking of two hands was performed, with the hands moving both separately and moving together. This is limited in its present state because it is not able to distinguish between the two hands if they are in close proximity to one another or touching each other.

The realisation that the trajectory of a single-handed gesture could be modelled as an aperiodic waveform instigated an investigation into how the harmonics of the Fourier analysis of the trajectory data could characterise the trajectory. The transformation of data from the time-domain to the frequency-domain had further advantages. Appropriate normalisation and removal of the d.c. component, of the frequency data enabled the harmonic description of the gesture to be position and scale independent. In addition the gesture duration was made time invariant by the application of multi-

rate methods to the time-domain data. The multi-rate technique changed the number of samples in the gesture to a constant number, thus allowing the same number of harmonics to be generated for all gestures. This allowed comparisons to be made based on the normalised period or frequency. The original objective of normalising to a length of sixty-four was based on the fact that most simple gestures lasted for less than two seconds. At twenty-five or thirty frames per second the multi-rate look-up table could work with most gesture lengths. This assumption was indeed true for many gestures, but with some gestures the duration of the gesture was two to three times longer than anticipated. It was shown that sub-sampling could be employed with negligible error in the frequency analysis, because the most important frequency content about a gesture was contained in just the first few low order harmonics.

Further analysis of the frequency data showed that all characteristics of the gesture's trajectory could be explained. Analysis of actual and pseudo trajectories showed that the phase of each harmonic was made from two components. These two components consisted of phase changes in the time domain and an angular or orientation component in the spatial domain. The time domain data typically resulted from a truncation of the waveform (start and finish coordinates do not coincide) and so phase changes compensate for the abrupt change in the waveform. The second component is the phase due to the orientation of the harmonic in the spatial or appearance–based domain. The first harmonic orientation angle is normally directly related to the coincident start/finish spatial coordinates and the top of the trajectory. The orientation angle was observed to be invariant to truncation errors, giving valuable insight into the characteristic of the gesture. For instance, the second harmonic orientation angle showed the amount of curvature there was in the appearance-based spatial view of the gesture, whereas the third harmonic orientation angle gave insight into the time-domain characteristics of the gesture. Further understanding of the structure of the gesture was to be found by studying the positive and negative sequence components of each harmonic. If each positive and negative sequence magnitude component were equal then the appearance-based view of the trajectory was a single line: the action and retraction path of the gesture was the same. However, when the positive and negative sequence magnitudes are different, the action and retraction paths are different. In the time-domain, each harmonic can be visualised as an 'elliptical corkscrew', with an ellipse being traced in the spatial domain with the spiral nature being seen in the time-domain. The number of rotations of the 'elliptical corkscrew' is in proportion to the number of harmonics. The direction of rotation of the 'elliptical corkscrew' is either anti-clockwise or clockwise depending if the magnitude of the positive sequence or negative sequence component is the larger. A gesture trajectory can thus be modelled as an infinite series of harmonics, each harmonic taking the form of an 'elliptical-corkscrew'.

The study of the frequency components can describe the characteristic of a gesture in detail, but it is useful to have a recognition system that can easily categorise a gesture. The advantage of using the PNN based on a RBF network, rather than some other method is because it avoids the need for extensive training, especially when calculating with sparse data. The PNN system is based on choosing 'target' or representative gestures followed by categorising the unknown gesture by its closeness to one of the targets. The PNN technique is similar to the 'least squares' techniques, but differs in the way it compares components of the input vector individually rather than globally, as in the least squares technique. The PNN is also

far less sensitive to outliers than the 'least squares' technique. The input to the PNN consisted of a vector that represented the harmonics of the magnitude and angle of the orientation harmonics that characterise a gesture. Each harmonic component was represented as the two parts of a complex number so as to avoid the discontinuity problem when using phase components straddle the 0°/360° boundary. For gestures which are very distinctive, as in pointing gestures, the first orientation angle indicates clear differences in direction of the gesture. As a result target gestures were relatively easy to estimate. In this case a typical gesture can be used as a target gesture. Due to the relatively large difference in orientation angle the system is quite insensitive to differences in gesturer because the orientation angle is so distinct. However, in other gesture situations a more reliable technique is required to derive target gestures.

The attractiveness of the PNN is that it can take multidimensional inputs and each dimension has its own scale factor automatically incorporated into the network, which avoids any scaling or any undue bias by unusual data values. Perhaps the main disadvantage is that all inputs are categorised by the nearness to the target gestures. There is no output that indicates that the target is not recognised. A possible solution to this would be to extend the number targets to then class them as main target misses.

Target gestures can be found by using clustering techniques. It had been found that using just a few low frequency harmonics the gesture trajectory could be reconstructed with minimal error. Hence, for clustering technique it was decided to use just the first three harmonics. The clustering technique showed that there are two important parameters to consider in clustering data. One is the distance metric and the other, is the linkage method. Dendrogram diagrams were useful for showing how the data was clustered. The 'ward' linkage method closely matched manual/visual methods. A technique was devised to investigate possible clustering based on clusters of the second and third harmonics data. This data was used to form target gestures with a PNN to which unknown gestures were categorised.

The combination of clustering technique with PNN was used to investigate inter-class and inter-class variations. These results compared very favourably with classification seen visually. The most significant find was that visual observations had identified that there were four subgroups to the particular gesture, albeit the fourth was rather large. It was very difficult to identify any common trait in the group. The cluster method effectively divided the gestures into five subgroups with much similarity to the visual technique.

Further insight into gestures came from the use of gesture stimuli. In this work the individuality of each person's gesturing became apparent. The results of this work showed the strong detection of beats in the gesture in addition to the unexpected duration of the gesture compared with pointing gestures.

8.2. Conclusions

The pivotal realisation of this thesis is that single-handed gestures can be modelled as an aperiodic waveform. Associated with developing this modelling is the realisation that the frequency analysis reveals important characteristics about the nature of the trajectory. The Fourier analysis of 1D waveforms can be analysed as an infinite

series of harmonics, usually of diminishing amplitude with frequency. In the analysis of gesture trajectories the movement of the hand is considered as a point moving in 2D space and sampled in the time domain. Exponential equations explain the characteristic of each harmonic in terms of positive and negative sequence components. The relative amplitudes of the positive and negative components describe ellipse structures in the spatial domain or appearance-based view. In the 2DT domain the ellipses are visualised as 'elliptical' corkscrews. The major axis of each ellipse (appearance-based view of each harmonic) also specified a unique orientation angle of each harmonic which allows for characterisation of the trajectory to be made. An additional benefit of using Fourier analysis techniques is that through normalisation techniques frequency data is made invariant to scale and translation effects of gesturing from sequence to sequence.

Experiments with 'pointing' gestures showed that there was close approximation between spatial position of the hand and the orientation angle of the first harmonic. The first orientation angle is basically formed from the spatial position at the start of the trajectory with the spatial position at the top of the trajectory. This correlation was shown for twenty of the twenty-one people repeating the same gesture because of the clearly defined start/stop and intermediate locations of the trajectory on the human body. It was also shown that for this latter type of gesture, the gesture trajectory could be synthesised to a good approximation from the first three harmonic components. The interpretation of the harmonic components confirmed the different movement primitives that had been observed by Gibet (2001) in sign language

By contrast the experiments with 'gesture stimuli' recorded oscillatory components in the trajectories that the Fourier analysis technique was ably equipped to capture. The oscillatory nature formed from gesturing tends to be constrained to whole cycles with very little spectral leakage as the hand is forced to return to near the starting coordinates of the gesture by gravity. The oscillatory nature of gestures may be considered as occurring during the stroke phase or after the stroke phase of a gesture. This suggests that the definition of a 'tri-phasic' gesture should be extended. The experimentation with gesture stimuli also showed that the occurrence of beats in gesture stimuli is more prevalent than in sign language.

In order to analyse gestures techniques, methods were developed to capture gesture trajectories and to normalise gesture lengths because of the variability in time of gesturer and gestures. Fusing of skin-colour cues with motion cues produced skin-coloured and motion objects. Rank ordering of these objects by size worked effectively at locating the dominant hand by the most significant object. It also avoided the need to adjust threshold values from sequence to sequence. However, in some sequences, colour or motion segmentation was difficult due to small amounts of motion or insufficient lighting. Furthermore, the first significant object was not suitable for tracking the dominant hand. The development of an object selection algorithm allowed for better tracking of the hand to occur. This led to extending the technique to the tracking of both hands and the tracking of a hand where there was movement from more than one person in an image sequence. The cascading of two interpolation and decimation functions allowed a range of gesture lengths to be normalised to the target length of sixty-four, with minimal error. These minor errors were seen as small phase shifts in the time domain which did not affect the characterisation (orientation angle) of the harmonics.

For the non-oscillatory trajectories the frequency analysis was not dependent on tracking the centroid of the hand shape. The skin-coloured and motion objects, especially in poor segmenting conditions could fragment occurring at various positions in the hand region. It was also found that only low-order frequency components were necessary to simulate the gesture trajectory as the variability of hand coordinates can be considered as high frequency noise.

However, gesture stimuli trajectories tested the system in a number of ways. Firstly, the trajectory length was generally double and in one case three times the length of pointing gestures it exceeded the sixty-four sample length target. This resulted in sub-sampling of the data reducing the highest frequencies recordable. Secondly, some of the oscillatory movements of the gesturer were of low amplitude that approached the amplitude range due to the variability of the coordinates of the gesture object on the hand region. In these cases the reliability of the tracked data or subsequent frequency analysis could not be relied upon as seen by the amplitudes of the higher order harmonics not 'tailing-off', as is the usual case with the harmonics of a Fourier series.

For those gesturing actions that have clearly defined spatial coordinates it was shown that intra-class and inter-class variations could be extracted from the data. The use of just the three lowest order harmonics, with the clustering technique and the PNN, could identify variations and sub-groupings in the gesturing activity. The gesture stimuli also showed the widely differing and idiosyncratic response of each gesturer, because of the lack of defined or imposed spatial coordinates on the gesture. This revealed that gesturers performed their own stylised response due to habit or their preconception of what was expected of them. This gave a unique characteristic to their individual gesture giving a potential feature for use in non invasive person recognition.

A by-product of the gesture recognition research is an area that another discipline may wish to research regarding human reactions. In the PET's sequences one of the gesturers had a 'red face' which was a different colour to his hands or the other gesturers' head and hands colour. This reaction is generally accepted as a consequence to some case of stressful or emotional situation. The other observation was the apparent involuntary rising of the non-dominant hand which accompanied the gesturer laughing. It was also noted that in the gesture stimuli set of experiments that many of the single handed gestures could not be undertaken without a movement of the non-dominant hand. These latter observations relate to the work that is now being undertaken on 'Affective Computing' concerning the emotional state of the gesturer.

8.3. Future research areas

The inspiration for one of the gesture experiments came from the work undertaken in avatar design. The use of 'avatars' illustrates how the analysis of human movement can be used to make animations very life-like. The avatar is defined from the position of all the joints when placed in some standard pose. Information is also required of each joint, whether it operates as a hinge or a ball and socket, or whether there are any limits of movement. The possibility of using reverse kinematics to feed hand coordinate information back from the analysis of actual hand movements from the

IFFT of harmonics data is a possibility that might reduce design times as well as giving more life-like gesturing.

The interesting feature of the control of avatars is in its use of quaternion mathematics. Quaternion mathematics are used for robot control because the problem of singularity or 'gimbal lock' is not encountered; they are more compact than conventional rotational matrix transformations and point to point interpolation is smoother. The 0°/360° discontinuity problem has occurred twice in this investigation; once in modelling Hue; secondly using phase information for the PNN network. Quaternion mathematics could be investigated for the modelling of Hue and other discontinuity situations. It is worth noting that Sanguine (2000) has used hyper-complex numbers or quaternions in the design of vector filters for colour images.

The generation of skin-colour and motion objects proved to be a robust method of detecting the hand position in a range of complex environments and even with poor lighting conditions. Recent work on the quality of an image has been undertaken by Luo (2005) into objective image measurements. These measurements could be useful in determining the ability of various techniques to work in poor conditions especially as it has been recognised that colour segmentation can fail or is difficult in some image sequences. Ranking the gesture objects into order, particularly based on area, increases the likelihood of the most significant object being related to the dominant hand. This reduced the complexity of hand tracking. The algorithm that determined what object to track could contend with various amounts of noise, but its performance deteriorated with poor lighting conditions. As a result it was not always able to track completely without a small amount of intervention. Further work is required to make the system work unattended with some method or system to segment the gestures automatically. Some additional criteria should be applied to the tracking criteria to improve performance, for example, gradient direction and inflection. Extending the tracking algorithm to track more than the first two most significant objects widens the scope: to track more accurately: to track both hands: to track more than one person simultaneously. Additional criteria would be needed for the algorithm track hand crossing or during occlusions.

The properties of skin colour could be investigated further for the detection of skin diseases. The work of Angelopoulou (2001), for example, showed that skin colour had a unique characteristic due to haemoglobin absorption at certain wavelengths. This property could be the focus for the development of a theory and method to consider whether haemoglobin absorption changed with the skin complaint. This could also assess if it could be detected in special lighting conditions or with a special colour model.

Research by other disciplines may be interested in how gesturing is affected by emotions. It has already been observed how skin-colour can change and the involuntary way the non-dominant hand and head are used in gesturing. The gesture stimuli experiments should be repeated but in a better lighting environment to aid segmentation and the tracking of relatively small hand movements. Working in a similar area, Ong and Ranganath (2005) have suggested future directions for gesture research beyond lexical meaning of automatic sign language. In particular they have proposed moving toward a true test of sign recognition systems that deal with natural signing by native signers.

There is the potential for further experimentation with a range of different applications of movements or activities associated with or similar to gesturing. The use of the PNN lends itself to having more than one input, so the input of two cameras could easily be used with this system. It appears that this technique might have a possible application in diagnosing medical conditions relating to the dexterity of the limbs; with any repetitive behaviour; or any manipulative activity. Some types of gesture or gesture stimuli might be ideal for the basis of a recognition system for a non-invasive security application.

From hypothesis to conclusion it has become clear that the gesture trajectory can be modelled as an aperiodic waveform in 2DT space. The analysis using positive and negative sequence components shows that the complete detail and characteristics of the trajectory can be realised by the properties of an infinite series of harmonic components. The investigation of gesture trajectories revealed the 'elliptical corkscrew' structure of each harmonic at different relative orientation angle to each other. However, this experimentation assumed that the camera for recording the image sequences was at right angles to the spatial domain. Further research should investigate the affect of camera view/projection on the phase of the harmonics, particularly the orientation angles. Additionally, depth movement of a point in the spatial domain is not recorded by this technique and so further research could be instigated to develop equations explaining the motion of a point in three dimensions and time (3DT).

This thesis has initiated the representation of gesture as an aperiodic waveform. This waveform representation has been analysed, characterised and classified, by the harmonic content using Fourier analysis technique. This is a new perspective on gesture analysis. The insight gained by this revelation, from hypothesis through experimentation and analysis to conclusions, must be considered as a fundamental foundation and catalyst for further research that could capitalise on the potential development of, and application of, a well grounded theory.

References

Agfa (1994), An *Introduction to Digital Scanning*, Digital Colour Prepress, Vol. 4.

Alon J., Sclaroff S., Kollios G. and Pavlovic V. (2003) Discovering Clusters in Motion Time-Series Data, *Proc. IEEE, CVPR*

Anderberg, M.R. (1973) *Cluster Analysis for Applications*. New York: Academic Press.

Angelopoulou, E. (2001), The uniqueness of the color of human skin, *SPIE's International Technical Group Newsletter*, June 2001, 5 and 9.

Austen, J. (1813), Pride *and Prejudice*, London: Penguin Books 1972, 184

Argyle, M. (1975), *Bodily Communications*. 2nd ed. London: Methuen & Co. Ltd.

Baeza-Yates, R. A. (1992), Introduction to data structures and algorithms related to information retrieval. *Information Retrieval: Data Structures and Algorithms*, Prentice-Hall, 13-27

Baudel, T. and Baudouin-Lafon, M. (1993) Charade: remote Control of Objects Using Free-Hand Gestures. *Comm. ACM*, **36,** no. 7, 28-35.

Bauer, B. Kraiss, K (2001), Toward an Automatic Sign Language Recognition System using Subunits, *International Gesture Workshop- GW 2001 London UK*, 64-75.

Berthold, M. (1994) A Time Delay Radial Basis Function Network for Phoneme Recognition. *Proceedings of the International Conference on Neural Networks*, Orlando California: Intel Corporation.

Bishop, C. M. (1995) *Neural Networks for Pattern Recognition*, Oxford University Press.

Bissell, C.C. and Chapman, D.A. (1992) *Digital Signal Transmission*, Cambridge

Black, J. Ellis, T. and Rosin, P. (2003) *A Novel Method for Video Tracking Performance Evaluation*, IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 125-132.

Bobick, A. F. and Davis, J. W. (2001) The recognition of Human Movement Using Temporal Templates, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **23**, 3, 257-267

Bobick, A. (1997) Movement, Activity, and Action: The Role of Knowledge in the Perception of Movement. *Philosophical Trans. Royal Soc. London*, 352, 1257-1265

Bobick, A. F. and Wilson A. D. (1997), A State-based Approach to the Representation and Recognition of Gesture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **19,** No 12, 1325-1337

Bobick, A. F. and Wilson A. D. (1995), A State-based Technique for the Summarization and Recognition of Gesture, *Proceedings of IEEE International Symposium on Computer Vision*, 382-388.

Bregler C. (1997) Learning and recognizing human dynamics in video sequences, *CPVR*, 568-574

Broomhead, D.S. and Lowe, D. (1988) Multivariable functional interpolation and adaptive networks, *Complex Systems*, **2**, 321-355

Cai, J. and Goshtasby, A. (1999) Detecting human faces in color images, *Image and Vision Computing*, **18,** 63-75.

Cannon, W. B. (1927) The James-Lange theory of emotion: A critical examination and an alternative theory. *American Journal of Psychology*; 39:10-124, [cited http://changingminds.org/explanations/theories/academic_references.htm#Cannon%20(1927) [accessed 15/12/2004]

Chen, F. S., Fu, C-M. and Huang, C. L. (2003) Hand gesture recognition using a real-time tracking method and hidden Markov models, *Image and Vision Computing*. **21**, 745-758.

Chen, C. and Chiang, S.P. (1997) Detection of human faces in colour images, IEE *Proc.-Vis. Image Signal Process.* **144,** No. 6, 384-388.

Chen, Q. Wu, H. and Yachida, M. (1995) Face detection by fuzzy pattern matching. *Proceedings of the Fifth International Conference on Computer Vision*, 591-596.

Corballis, M. C. (2003) Form mouth to hand: Gesture, speech, and the evolution of right-handedness. *Behavioural and Brian Sciences*, **26**, 199-260.

Dai Y. and Nakano, Y. (1995) Extraction of Facial Images from Complex Background Using Color Information and SGLD Matrices, *International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 238-242.

Dai Y. and Nakano, Y. (1996) Face-texture model based on SGLD and its applications in face detection in a colour scene, *Pattern Recognition,* **29**(6), 1996, 1007-1017, cited in J. Cai, A. Gostasby, Detecting human faces in color images, Image and Vision Computing **18** (1999), 63-75

Darrell T. and Pentland A. P. (1993) Space-time gestures. *Proc. Comp. Vis. And Pattern Rec.*, 335-340.

Daugman, J. (1997) Face and Gesture Recognition: Overview. *Pattern Analysis and Machine Intelligence*, **19,** No7, 675-676.

Davies E. R. (1997), *Machine Vision: Theory, Algorithms, Practicalities*, 2nd Ed., Academic Press, pp. 103-114.

D'Cunha I and Alvertos, N. (1994), Blob tracking technique to Range Image Segmentation, *Proceedings of SPIE (The International Society for Optical Engineering),* **2353**, 216-223.

Davis J.W. and Bobick A.F. (1996) The Representation and Recognition of Human Movement Using Temporal Templates. *Proceedings of Automatic Face and Gesture Recognition, Killington, VT*, 928-935.

Duba,http://www.cs.princeton.edu/courses/archive/fall05/cos436/Duda/PR_Mahal/M_metric [Accessed June 2006]

Ekman, P. and Friesen, W. V. (1969) The repertoire of nonverbal behaviour: categories, origins, usage, and coding, *Semiotica*, **1**, 49-98.

Elliott, R. J. Glauert, J.R.W. Kennaway J.R. and Marshall, I. (2000) The development of language processing support for the ViSiCAST project. *Proc. 4th International ACM (Conference on Assistive Technologies),* 101-108.

Ellis, T. (2002) Performance Metrics and Methods for Tracking in Surveillance, *Proceedings 3rd IEEE Int. Workshop on PETS.*

FGNet, http://www.cvmt.dk/~fgnet/apps.html[Accessed December 2004]

FGNet, http://www.fg-net.org with additional support provided by the Swiss National Centre of Competence in Research (NCCR) on Interactive MultiModal Information Management (IM)2.

Freeman, W. T. and Roth, M. (1995) Orientation Histograms for Hand Gesture Recognition, Proc., *International Workshop on Automatic Face and Gesture Recognition, Zurich, Switzerland*

Gavrila, D.M. and Davis, L.S. (1996), 3-D model-based tracking of humans in action: a multi-view approach. *In Proc. IEEE Conf. On Computer Vision and Pattern Recognition.*

Gavrila D M, (1999), The Visual Analysis of Human Movement: A Survey, *Computer Vision and Image Understanding*, **73**, No.1, 82-98.

Gibet, S. Lebourque, T. and Marteeau, P-F. (2001) High-level Specification and Animation of Communicative Gestures, *Journal of Visual Languages and Computing* **12**, 657-687.

Gonclaves, L. Di Bernardo, E. Ursella, E. and Perona, P. (1996) Monocular Tracking of the Human Arm in 3D, *In Proc. IEEE Int. Conf. Computer Vision and Pattern Recognition.*

Goldin-Meado, S. and Mylander, C. (1998), Spontaneous sign systems created by deaf children in two cultures, *Nature*, **39**, 279-281.

Gong S., Ng J., Sherrah J.,(1998) "On the Semantics of Visual Behaviour, Structured Events and Trajectories of Human Action", *Image and Vision Computing*, 20, Issue 12, 873-888.

Gong, S. McKenna, S.J. and Psarrou A. (2000) *Dynamic Vision – From Images to Face Recognition,* Imperial College Press.

Gonzalez, R.C. & Woods, R.E. (1992) *Digital Image Processing*, Addison-Wesley, pp. 229-235.

Grove, N. and Walker, M. (1990) The Makaton Vocabulary: Using Manual Signs and Graphic Symbols to Develop Interpersonal Communications. *Augmentative and Alternative Communication*, Williams & Wilkins.

Hagan, M.T. (1996) *Neural Network Design*. Boston: PWS Publishing Company.

Hanbury, A. and Serra, J. (2001) Mathematical Morphology in the HLS Colour Space, *BMVC*, 451-460.

Harding, P. (1999), Investigations into Skin Colour as a Feature for Detecting Hand Gestures. *Technology Letters*, **3**, 1, 34-43.

Harding P R G and Ellis T J, (2003), An Improved Skin Colour Segmentation Technique for Hand Gesture Tracking, *Poster at International Gesture Workshop, Genoa.*

Hardy, T. (1891) Tess of the D'Urbervilles, Penguin: London, 1978.

Haykin, S. and Van Veen, B. (1998) *Signals and Systems*, New York: Wiley

Heap, T. and Hogg, T. (1998), Wormholes in Shape space: Tracking through Discontinuous Changes in Shape, *Proc. ICCV*, 344-349.

Holte, M. B. and Storring, M. (2002) Documentation of Pointing and command gestures under mixed illumination conditions: video sequence database. http://www.cvmt.dk/~fgnet/Pointing04/, [Accessed 12/11/2003]

Huang T. S. and Pavlovic, V.I. (1995), Hand Gesture Modelling, Analysis, and Synthesis, *International Workshop on Face- and Gesture-Recognition*, 73-78.

HUMOSIM, The University of Michigan Laboratory for Human Motion Simulation http://www.engin.umich.edu/dept/HUMOSIM, [Accessed 28/01/2004]

Howell, A.J. Sage, K. and Buxton, H. (2003), Developing Task-Specific RBF Hand Gesture Recognition, *Cognitive Science Research Papers, ISSN*, 1350-3162.

Howell A. J. and Buxton H. S. (1998) Learning Gestures for Visually Mediated Interaction, *Proc. BMVC*, 508-517.

Howell A. J. and Buxton H. S. (1995) Invariance in Radial Basis Function Neural Networks in Human Face Classification, International Workshop on Automatic Face- and Gesture- Recognition, 221-226.

Ifeachor, E.C. Jervis, B.W. (1993) Digital Signal Processing – A Practical Approach, Addison Wesley.

Isard M. and Blake A. (1998) Condensation – Conditional Density Propagation for Visual Tracking, *Int. Journal of Computer Vision*, **28**, 1, 5-28.

Jain, A. K. and Mao, J. (1996) Artificial Neural Networks: A Tutorial. *Computer*. 0018-9162, 31–44.

Jain A. K., Murty M. N. and Flynn P.J. (1999) Data Clustering: A review, ACM Computing Surveys, **31**, 264-323

Jain, R. Kasturi, R. and Schunch, B. G. (1995) *Machine Vision*, McGraw-Hill, 408-420.

Jackson, R. MacDonald, L. and Freeman, K. (1994) *Computer Generated Color (A Practical Guide to Presentation and Display*), Wiley.

Johansson, G. (1973) Visual Perception of Biological Motion and a Model for its Analysis, Perception and Psychophysics. **14**, 2, 201-211.

KaewTraKulPong, P. and Bowden, R. (2001) An Adaptive Visual System for Tracking Low Resolution Colour Targets, *BMVC 2001*, 243-252.

Kangas, J. (1990), Time-delayed self-organizing maps, *Proceedings of the International joint Conference on Neural Networks*, 331-336

Kelly, S.D. Barr, D.J. Breckinridge-Church, R. and Lynch K. (1999) Offering a Hand to Pragmatic Understanding: The Role of Speech and Gesture in Comprehension and Memory, *Journal of Memory and Language*, **40**, 577-592.

Kendon, A. (1986) Current Issues in the Study of Gesture, *The Biological Foundations of Gestures: Motor and Semiotic Aspects*, J L Nespoulous, J.L. Peron P. and Lecours, A. R. eds., Lawrence Erlbaum Assoc.,23-47.

Kennaway, R. (2001) Synthetic Animation of Deaf Signing Gestures, *International Gesture Workshop, GW 2001*, 146-157.

Kennaway, R. (2003) Experience with and requirements for a Gesture Description Language for synthetic Animation, *5th International Gesture Workshop, GW2003*, 300-311.

Kennaway, R. (2004) 'Take Mug' avatar animation sequence, personal communication.

King B. (1967), *Step-wise clustering procedures*, J. Am. Stats. Assoc. **69**, 86-101

Kjeldsen, R. and Kender, J. (1996) Finding Skin in Color Images, *Proceedings of Automatic Face and Gesture Recognition*, 312-317.

Kobayashi, T. and Haruyama, S. (1997) Partly Hidden Markov Models and its Application to Gesture Recognition, *Proc. Int'l Conf. Acoustics, Speech and Signal Processing*, **4**, 3081-3084

Kohler, M. (2001), Vision Based Hand Gesture Recognition Systems, Computer graphics, University of Dortmund. http://ls7-www.cs.uni-dortmund.de/research/gesture/vbgr-table.html [Accessed 26/3/2001]

Kohonen T. (1988) *Self-Organising and Associative Memory*. New York: Springer-Verlag

Koski, L. Iacoboni, M and Mazziotta, J. C. (2002) Deconstructing apraxia: understanding disorders of intentional movement after stroke. *Current Opinion in Neurology*, **15,** 71-77.

Kraniauskas P. (1992) *Transforms in Signals and Systems*, Addison-Wesley

Kuhl F. P. and. Giardina C.R. (1982) Elliptic Fourier Features of a Closed Contour, *Computer Graphics and Image Processing*, **18**, 236-258.

Lang, K. and Hilton, G. (1988*) A time-delay neural network architecture for speech recognition*. Tech. Report CMU-CS-88-152, Carnegie-Mellon University.

Lapedes, A., and Farber, R. (1987) *Non-linear signal processing using neural networks: prediction and system modelling*, Technical Report LA-UR-87-2662, Los Alamos National Laboratory.

Lee, C. H. Kim, J. S. Park, K. H. (1996) Automatic Face Location in a Complex Background using Motion and Color Information, *Pattern Recognition*, **29**, 11, 1877-1889.

Lockton R. and Fitzgibbon A. W. (2002) Real-time gesture recognition using deterministic boosting. *BMVC 2002*, 817-829.

Lin, C.S. and Hwang, C.L. (1987) New Forms of Shape Invariants from Elliptic Fourier Descriptors. *Pattern Recognition*, **20,** 5, 535-545.

Lin, C.S. and Jungthirapanich, C. (1990) Invariants of Three-Dimensional Contours. *Pattern Recognition*, **23**, 8, 833-842.

Lin, D., J. Dayhoff and P.A. Ligomenides. (1992) Adaptive time-delay neural networks for temporal correlation and prediction, *SPIE Intelligent Robots and Computer Vision XI: Biological Neural Net and 3-D Methods*, Boston, 170-181.

Lin, D. Dayhoff, J. and Ligomenides, P.A. (1995) Trajectory Production with the Adaptive Time-Delay Neural Network. *Neural Networks*, Pergamon Press, **8**, 3, 447-461.

Lincoln, S. J. Cox, M. and Nakisa, M. (2001) The development and evaluation of a speech to sign translation system to assist transactions. *Int. Journal of Human-computer Studies.*

Linde, Y., Buzo A. and Gray R. M. (1980) An Algorithm for Vector Quantizer Design, *IEEE Trans. Comm.*, 28(1): 84-95

Luo, G. (2005), *A novel technique of image quality objective measurement by wavelet analysis throughout the spatial frequency range.* Proc SPIE-IS&T Electronic Imaging, SPIE Vol. 5668, 173-184

Lynn, P. A. (1994), An Introduction to the Analysis and Processing of Signals, 2nd Ed., Macmillan.

Mammen, J.P., Chaudhuri, S. and T Agrawal, T. (2001) Simultaneous Tracking of Both Hands by Estimation of Erroneous Observations, *Proceedings of the British Machine Vision Conference*, 83-92.

Masters, T. (1994) *Signal and image processing with Neural Networks*, John Wiley and Sons.

Matlab V5, The Math Works, Inc, 1997

McCulloch, W.S. and Pitts W. (1943), A logical calculus of the ideas immanent in nervous activity, *Bulletin of Mathematical Biophysics*, **5**, 115-133.

McKenna, S., Gong, S. and Y Raja, Y. (1997) Face Recognition in Dynamic Scenes, *Proceedings of the British Machine Vision Conference*, 140-151.

McKenna, S., Gong, S. and Y Raja, Y. (1998), Modelling Facial Colour and Identity with Gaussian Mixtures, *Pattern Recognition*, **31**, 12, 1883-1998.

McKenna S. J. and Gong S. (1998) Gesture Recognition for Visually Mediated Interaction using Probabilistic Event Trajectories, *Proceedings of the British Machine Vision Conference*, 498-507.

McNeil, D. and E Levy, E. (1982) Conceptual Representation in Language Activity and Gesture, Speech, Place and Action, *Studies in Delixis and Related Topics*, J Jarvella & W Klein, eds. Wiley.

McQueen J. (1967) Some methods for classification and analysis of multivariate observations, *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 281-297.

MIT, Affective Computing Group http://affect.media.mit.edu/ [Accessed 20/9/2004]

Moon, T.K. and Stirling, W. C. (2000), *Mathematical Methods and Algorithms for Signal Processing*, Prentice Hall, New Jersey, 73.

Moody, J. and Darken C.J. (1989), Fast learning of locally-tuned processing units, *Neural Computation*, **1**, 281-294.

Morris D. (1967), *The Naked Ape*, Jonathan Cape London

Morris D. (1977), *Manwatching. A field-guide to human behaviour*, Jonathan Cape/Elsevier, London/Oxford.

Morris, D., Collett, P., Marsh, P. and O'Shaughnessy M. (1979), '*Gestures, Their Origins and Distributions*', Jonathan Cape Ltd.

Morrison, K. and McKenna, S. J. (2003), An Experimental Comparison of Trajectory-based and History-based representation for Gesture Recognition, *International Gesture Workshop 2003*, 152-163.

Mowbray, S.D. and Nixon, M.S. (2003) Automatic Gait Recognition via Fourier Descriptors of Deformable Objects, eds. Kittler, J. and Nixon, M. S., *Proceedings Audio Visual Biometric person Authentication (AVBPA 2003)*, 566-573.

Murtagh F. (1984) *A survey of recent advances in hierarchical clustering algorithms which use cluster centres*, Comput. J. 26, 354-359

Nagy, G. (1968) State of the art in pattern recognition, *Proc. IEEE* **56**, 836-862

Ng C. W. and Ranganath, S. (2002) Real-time gesture recognition system and application, *Image and Vision Computing*, **20**, Issue 13-14.

Ondaatye, M. (1993) '*The English Patient*', Picador 1993 (first published 1992), 38.

Ong, S and Ranganath, S. (2005) Automatic Sign Language Analysis: A Survey and the Future beyond Lexical Meaning, *IEEE PAMI*, **27**, 6, 873-890.

Orr, M. (1996) *Introduction to Radial Basis Function Networks*. Edinburgh: University of Edinburgh.

Oulu, The University of Oulu Physics-Based Face Database, http://www.ee.oulu.fi/research/imag/color/pbfd.html, [Accessed 20/1/2005]

Owens, F. J. (1993) *Signal Processing of Speech*, Macmillan.

Parzen, E. 1962, On estimation of a probability density function and mode, *Annals of Mathematical Statistics*, **33**, 1065-76.

Pavlovic, V. I. Sharma, R. and Huang, T. S. (1996) Gestural Interface to a Visual Computing Environment for Molecular Biologists, *Proceedings of Automatic Face and Gesture Recognition,* 30-35.

Pavlovic, V. I. Sharma, R. and Huang, T. S. (1997) Visual Interpretation of Hand Gestures for Human-Computer Interaction: A Review, *Pattern Analysis and Machine Intelligence*, **19**, 7, 677-695.

Perez, P., Hue, C., Vermaak, J., and Gangnet, M. (2002) Color-Based Probabilistic Tracking, *ECCV*, pp. 661-675.

Peterfreund, N. (1999) Robust Tracking of Position and Velocity with Kalman Snakes', *IEEE trans. PAMI,* **10**.(6), 564-569.

Pezeshkpour, F. Marshall, I. Elliott, R. and Bangham, J. A.(1999) Development of a Legible Deaf-Signing Virtual Human. *Proc. IEEE Conf. Multi-Media*, **1**, pp. 333-338.

Pichard, R. (2004) http://www.media.mit.edu/affect [Accessed 20/9/2004]

Picton P. (2000) *Neural Networks*, 2nd Ed., Palgrave, 104-109.

Pitas, I. (1993) *Digital Image Processing Algorithms*, Prentice Hall.

Poggio, T. and Girso, F. (1990) Regularization algorithms for learning that are equivalent to multiplayer networks, *Science*, **247**, 978-982.

Poynton C. (1997), Frequently Asked questions about Color, www.poynton.com, [Accessed 19/9/2004]

Prillwitz, S. Leven, R. Zienert, H. Hanke, T. Henning, J. et al. (1989) HamNoSys Version 2.0: Hamburg Notation System for Sign Languages – An Introductory Guide. International Studies on Sign Language and Communication of the Deaf, **5**, University of Hamburg, Version 3.0, http://www.sign-lang.uni-hamburg.de/Projects/HamNoSys.html [Accessed 17/1/2003]

Quek, F. K. H. (1994) Toward a Vision-Based Hand Gesture Interface, *Virtual Reality Software and Technology Conference.* 17-31.

Person E. and Fu, K. (1977) Shape Discrimination Using Fourier Descriptors', *IEEE Transactions on Systems, Man, and Cybernetics,* SMC-**7,** 3, 170-179.

PETS, ftp://pets.rdg.ac.uk/PETS-ICVS [Accessed 6/10/2002]

Rabiner, L R (1989) A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition, *Proceedings of the IEEE*, **77,** 2, 257-286.

Renals, S. (1989) Radial basis function network for speech pattern classification, *Electronic Letters*, **25**, 437-439.

Rigoll, G. Kosmala, A. and Eickeler, S. (1997) High Performance Real-Time Gesture Recognition using Hidden Markov Models, *Gesture and Sign Language in Human-Computer Interaction: International Gesture Workshop.*

RNIB (Royal National Institute for the Blind) http://www.rnib.org.uk/xpedio/groups/public/documents/visugate/public_movegest.hcsp last updated 12/10/04 18:25 [Assessed 14/11/2004]

Rossini, N. (2003) The Analysis of Gesture: Establishing a set of Parameters, *Gesture-Based Communication in Human-Computer Interaction, 5th International Gesture Workshop*, 124-131.

Sage, K., Howell, A.J. and Buxton, H. (2003) Developing Context Sensitive HMM Gesture Recognition, *International Gesture Workshop 2003*, 277-287.

Sangwine S. J. (2000) Colour in Image Processing, *Electronics & Communication Engineering,* IEE, 211-219.

Schiele, B. and Waibel, B. (1995) Gaze tracking Based on Face-Color, *International Workshop on Automatic Face- and Gesture-Recognition*, Zurich, 344-349.

Schraudolph, N. and Cummins, F. 2001. *Introduction to Neural Networks,* http://gened.emc.maricopa.edu/bio/bio181/BIOBK/BioBookNERV.html [Accessed 15/2/2001].

Sharma, G. and Trussell, H. J. (1997) Digital Color Imaging, *IEEE Transactions on*

Image Processing, **6**(7), 901-927.

Sherrah, J. and Gong, S. (2000), Gesture Recognition for Visually Mediated Interaction, http://nick.dcs.qmul.ac.uk/~sgg/gesture/ [Accessed 21/10/2003]

Sherrah, J. and Gong, S. (2001), Fusion of Perceptual Cues for Robust Tracking of Head Pose and Position, *Patter Recognition*, 34(8), 1565-1572.

Siebel, N.T. and Mabank, S. (2002) Fusion of Multiple Tracking Algorithms for Robust People Tracking, *ECCV*, 373-387.

Sigal, L. Sclaroff S. and Athitos, V. (2004) Skin Color-Based Video Segmentation under Time-Varying Illumination, *IEEE Trans. PAMI*, **26**, 7, 862-877.

Siple, P. (1978) *Linguistic and Psychological Properties of American Sign language: An Overview in Understanding Language Through Sign Language Research* (ed. P.Siple): Academic Press, New York, pp. 14.

Smith, A. R. (1978) Color Gamut Transform Pairs, *SIGGRAPG '78*, 12-19.

Sneath P.H. and Sokal R.R. (1973) *Numerical Taxonomy*, Freeman, London.

Sobottka, K. Pitas, I. (1996) Segmentation and tracking of faces in colour images, *Proceedings of the Second International Conference on Automatic Face and Gesture Recognition*, 236-241.

Starner, T.E. and Pentland, A. (1995) Visual Recognition of American Sign Language Using Hidden Markov Models, *Proc. First Int. Workshop Automatic Face and Gesture Recognition,* 189-194.

Stauffer C. and Grimson, W. E. L., (1999) Adaptive background mixture models for real-time tracking, Proc. *IEEE CVPR Conf.*

Stefanov N., Galata A. and Hubbold R. (2005) Real-time Hand Tracking with Variable-Length Markov Models of Behavior, *IEEE Int. Workshop on Vision for Human-Computer Interaction in conjunction with CVPR*

Sturman, D.J. Zeltzer, D. (1994) A Survey of Glove-Based Input, *IEEE Computer Graphics and Applications,* **14**, 30-39.

Terrillon, J.C. David M. and Akamatsu, S. (1998) Detection of Human Faces in Complex Scene Images by Use of a Skin Color Model and of Invariant Fourier-Mellin Moments, *Proceedings 14th International Conference on Pattern Recognition,* **II**, 1350-1356

Terrillon, J. C. and Akamatsu, S. (1999) Comparative Performance of Different Chrominance Spaces for Colour Segmentation and Detection of Human Faces in Complex Scene Images, *Proc. Vision Interface*, 180-187.

Vermaak, J. Perez, P. Gangnet, M. and Blake A. (2002) Towards Improved Observation Models for Visual Tracking: Selective Adaption, ECCV, 645-660.

Vernon, D. (1991) *Machine Vision*: *Automated Visual Inspection and Robot Vision*, Prentice Hall, pp. 99-100.

ViSiCAST, http://www.visicast.sys.uea.ac.uk/Public_pictures.html [Accessed 15/10/2004]

Waibel, A. Hanazawa, T. Hinton, G. Shikano, K. and Lang, K. (1989). Phoneme Recognition Using Time-Delay Neural Networks, *IEEE Trans. Acoustics, Speech, and Signal Processing*, **37**, 3, 328-339.

Wallace, T.P. and Mitchell, O.R. (1980) Analysis of Three-Dimensional Movement Using Fourier Descriptors. *IEEE Trans. Pattern Analysis and Machine Intelligence*, **2**, 6, 583-588.

Wallace T. P. and Wintz, P. A. (1980), An Efficient Three-Dimensional Aircraft Recognition Algorithm Using Normalised Fourier Descriptors, *Computer Graphics and Image Processing*, **13**, 99-126.

Ward J. H. (1963) *Hierarchical grouping to optimise an objective function*, J. Am. Stats. Assoc. **58**, 236-244

Weigend, A, and N. Gershenfeld, N. eds. (1993) Time *Series prediction: Forecasting the Future and Understanding the Past*, Addison-Wesley.

Werbos, P. (1974*) Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*, Ph.D. thesis, Harvard University, Cambridge, MA.

Wilson, A. D. and Bobick, A. (1995) Learning Visual Behaviour for Gesture Analysis, *Proceedings of IEEE International Symposium on Computer Vision*, 229-233.

Wilson, A. D. Bobick, A. and Cassell, J. (1996) Recovering the Temporal Structure of Natural Gesture, Proceedings *of Automatic Face and Gesture Recognition*. 66-71.

Woll, B. (2004), The City University, personal dialogue with Professor of Sign Language and Deaf Studies.

Wozniak, R. H., http://www.thoemmes.com/psych/james.thm. [accessed 15/12/2004]

Wren, C. R. Azarbayejani, A. Darrell, T. and Pentland A. P. (1997) Pfinder: Real-Time Tracking of the Human Body, *IEEE Transactions on Pattern Analysis and Machine Intelligence,* **19**, 7, 780-785.

Yamato, J. Ohya, J. Ishii, K. (1992) Recognizing Human Action in Time-Sequential Images using Hidden Markov Model, *Computer Vision and Pattern Recognition*, 379-385.

Yam C.Y. Nixon M. S., Carter J. N. (2002) Gait Recognition by Walking and Running: A Model-Based Approach, The 5^{th} *Asian Conference on Computer Vision*.

Yang M. H., Abuja, N. and Taub, M. (2002) Extracting of 2D Motion Trajectories and Its Application to Hand Gesture Recognition" *IEEE Trans. Pattern Analysis and Machine Intelligence*, **24**, 8, 1061-1074.

Zahn C. T. and Roskies, R. Z.(1972) Fourier Descriptors for Planed Closed Curves, IEE*E Transactions on Computers*, **21,** 3, 269-281.

Appendix I – Avatars

Avatar Design

Information about avatars is available from: -
http://viscastsys.uea.ac.uk/Public_pictures.html

Measurement of the position of body parts are shown in figure A1.1 from the avatar sequence 'motion_capature.avi'.

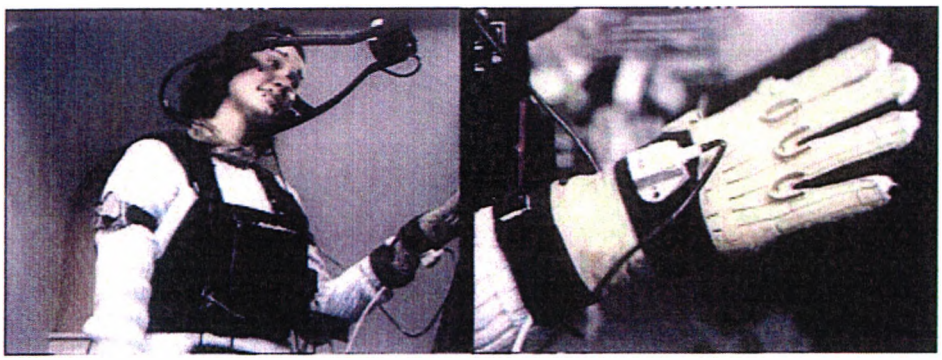


Figure A1.1 Measurement of Body Positions for Avatar Design

The stages of avatar realisation are shown in the avatar sequence 'tessa.avi'. A skeletal model is show in the first frame and an outline shape is shown in frame 295 of figure A1.2.

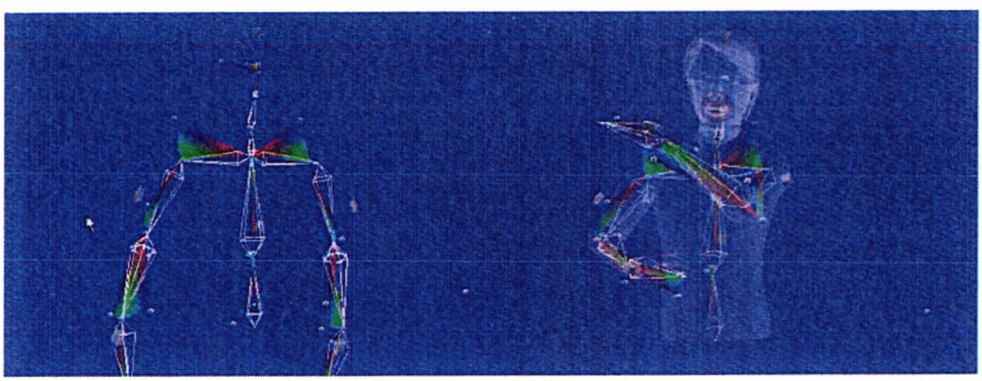


Figure A1.2 Skeletal (frame 1) and Outline Figure (frame 295) of avatar Tessa

The final clothed avatar is shown in frame 660 of figure A1.3.

Figure A1.3 Tessa the avatar (frame 660)

Figure A1.4 shows an image from the 'Take Mug' sequence that inspired a range of experiments (Chapter 7).

Figure A1.4 A frame from the 'Take Mug' sequence

Avatar Specification (Kennaway - 2001)

The synthesis of deaf signing animations is accomplished from a high-level description of signs in terms of the HamNoSys transcription system. The HamNoSys system defines many contact points on the body, such as positions at, above, below, left, or right of each facial element (eyes, nose, checks, etc.) and several positions on each finger and along the arms and torso. The total number of positions nameable in HamNoSys comes to some hundreds.

"Given a definition of the numerical coordinates of all the hand positions described by HamNoSys, we must determine angles of the arm joints which will place the hand in the desired position and orientation. This is the problem in 'inverse kinematics' (forward kinematics being the opposite and easier problem, of computing hand position and orientation from the arm joint angles)."

Quaternion rotational transformations are used for interpolation from one point to another. Quaternions offer several advantages over traditional transformations.

- Problem of singularity is not encountered.
  Singularity is also known as 'Gimbal Lock'.
- More compact than conventional transformations.
  Conventional rotational matrix is made up of 9 numbers.
- Quaternion rotation is made up of 4 numbers.
- Point to point interpolation is smoother

When a quaternion is used to represent rotations, the first three components are a vector parallel to the axis of rotation, and the length of that vector is the sine of half the rotation angle. The fourth component is the cosine of half the rotation angle. (Some authors put the fourth component first.) To calculate the composition of two rotations in quaternion form, the rotation of a vector by a quaternion, or the conversion between quaternion and rotation matrices, see mathematical sources for information.

To find where the wrist is, you need to combine all the translations and rotations at the joints up the hierarchy from the wrist to the root. You can either do that using quaternion directly, or convert them all to rotation matrices first and work with those.

Avatar 'Take-Mug' sequence

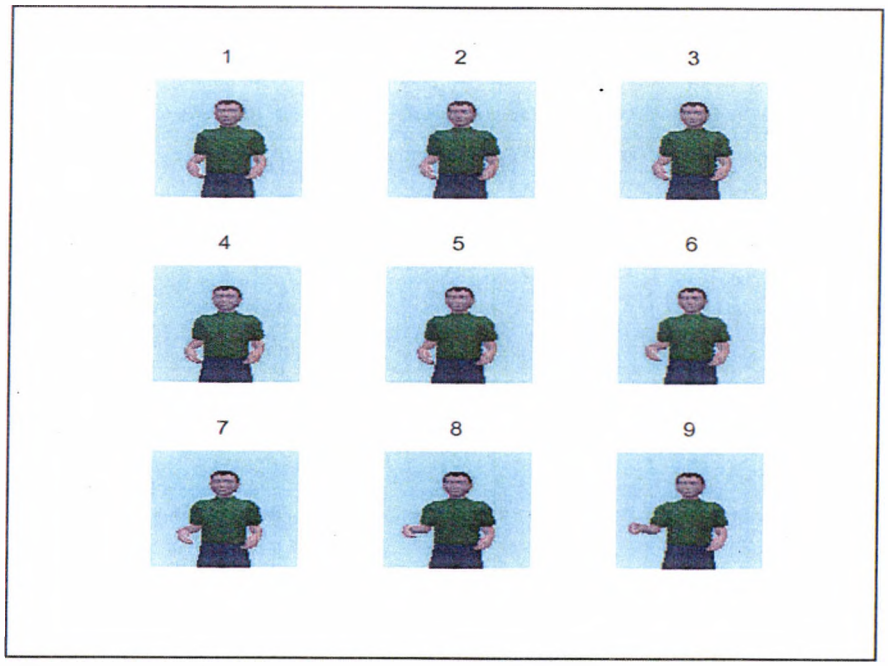


Figure A1.5 Frames 1 to 9 from the avatar 'Take Mug' sequence

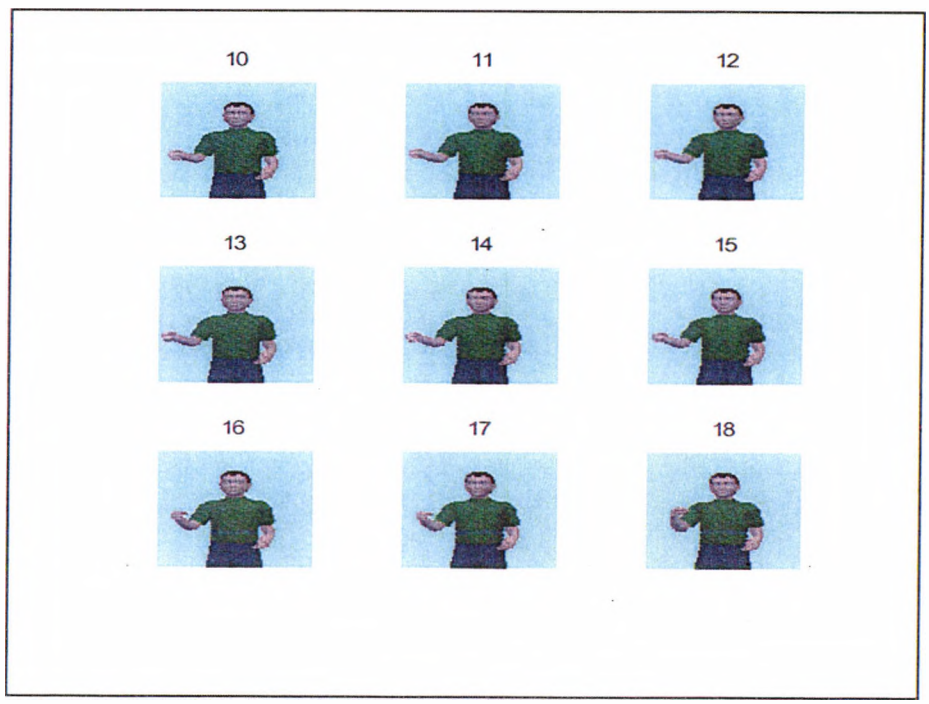


Figure A1.6 Frames 10 to 18 from the avatar 'Take Mug' sequence

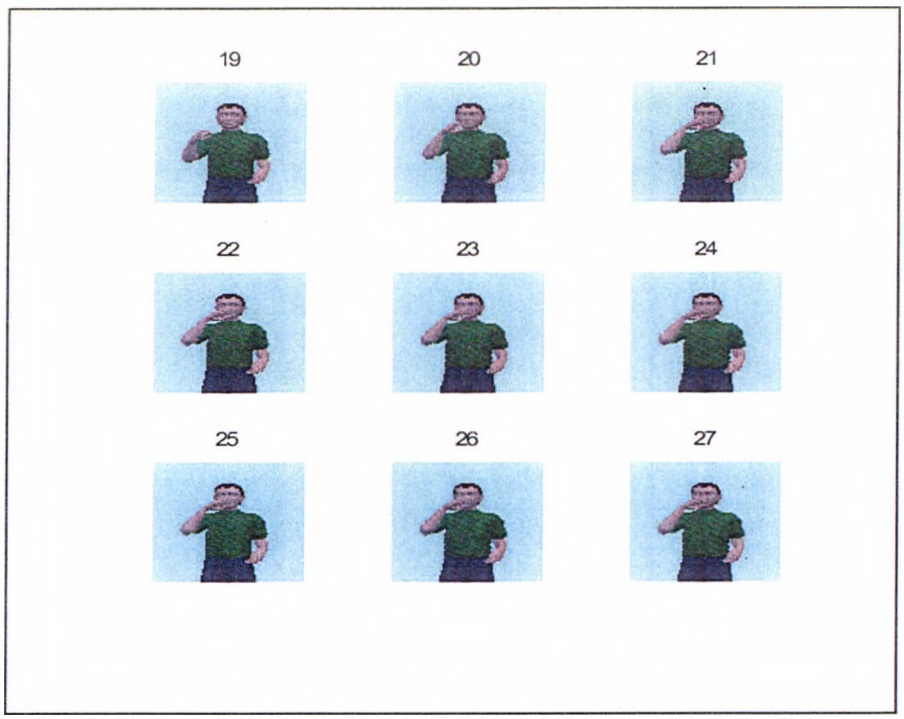


Figure A1.7 Frames 19 to 27 from the avatar 'Take Mug' sequence

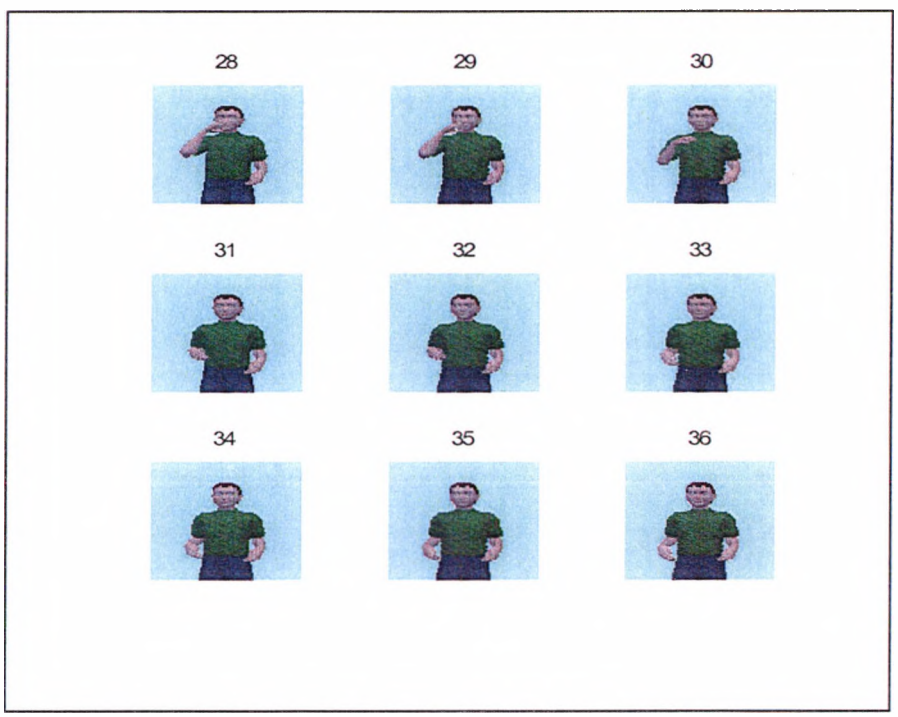


Figure A1.8 Frames 28 to 36 from the avatar 'Take Mug' sequence

Appendix II – Skin-Colour Variations due to Different Illuminants and White Balance Corrections

A Skin Colour Uniqueness

Angelopoulos (2001) identified that there is uniqueness to the colour of human skin. The reflectance of the back of the hand and the palm were measured from a diverse group of people from Caucasian, Asian and African descent. Measurements showed that that the overall percentage of light that was reflected from human skin increased with wavelength. Around the 575nm wavelength there was a specific shape that looks like the letter W (two dips with a bump in the middle) as shown in Figure 3.3. The uniqueness of this characteristic was confirmed by comparing the reflected light to that of a mannequin. The mannequin had been designed to be as life-like as possible but gave a distinctly different spectral characteristic. A biological explanation for this characteristic was forthcoming by observing the absorption spectrum of oxygenated haemoglobin. The absorption spectrum of haemoglobin exhibits the inverse W pattern (i.e. an M pattern) at almost identical wavelengths (i.e. 542nm, 560nm and 576nm respectively, as shown in Figures 3.4. The University of Oulu Physics-Based Face database (2001) has results of skin spectral reflectance characteristics. The reflectance of the cheek and forehead were measured and showed a similar dip in the spectral response between 520nm and 590nm, albeit the W form was not so clearly identifiable.

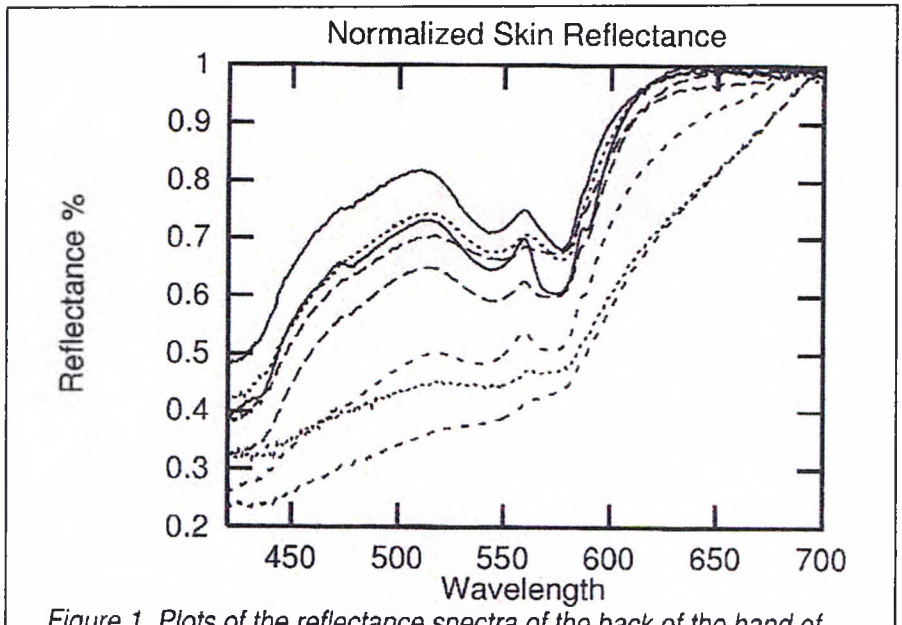


Figure 1. Plots of the reflectance spectra of the back of the hand of various subjects.

Figure A2.1 Plots of the reflectance spectra of the back of the hand of various subjects (Source: Angelopoulos, 2001)

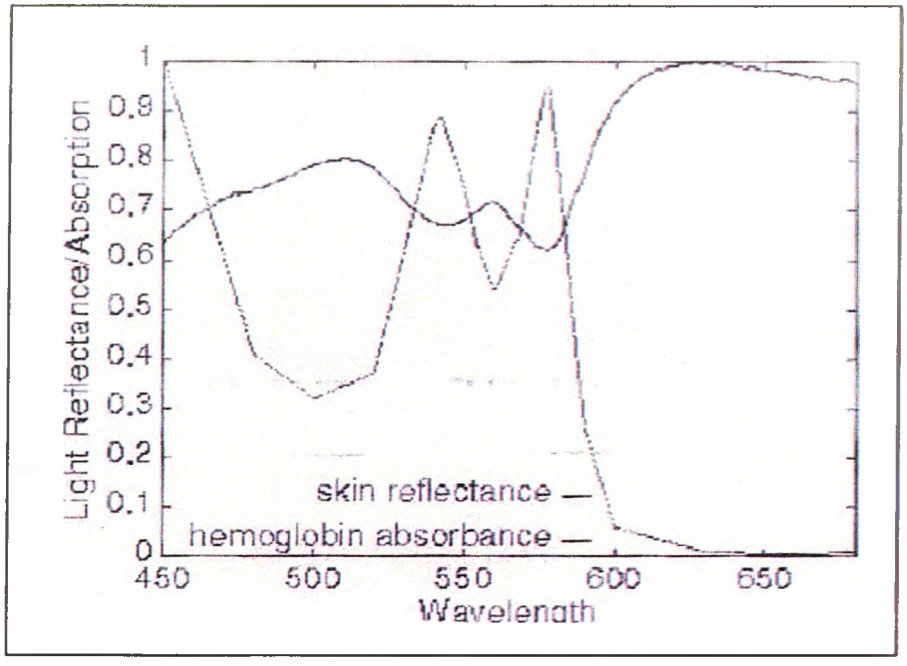


Figure A2.2 The reflectance spectrum of human skin compared with the absorption spectrum of oxygenated haemoglobin (Source: Angelopoulos, 2001)

B HSI and HSV Colour Models

HSI formulae:

The HSI space components are given as (Gonzalez and Woods, 1992): -

$$H = \cos^{-1}\frac{(2R-G-B)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}, \text{deg}\,rees$$

$$then, H = H/360$$

$$S = 1-\frac{3}{(R+G+B)}\min(R,G,B)$$

$$I = \frac{1}{3}(R+G+B)$$

HSV algorithm:

Given R, G, and B, each on domain [0,1]. Desired: The equivalent H, S, and V, each on range [0, 1].

1 V: = max(R, G, B);

2 Let X: = min(R, G, B);

3 S: = (V-X)/V; if S=0 return;

4 let r: = (V-R)/(V-X);

g: = (V-G)/(V-X);

b := (V-B)/(V-X);

5 If R = V then H: =(G=X then 5+b else 1-g);

If G = V then H: =(B=X then 1+r else 3-b);

else H: = (if R=X then 3+g else 5-r);

6 H: = H/6;

Remarks: H=0 is taken to be Red (G=B and R>B) by convention.

The Hue derivation (HSI model) is very intense on mathematical functions and can be a restricting factor when speed is a premium. A calculations comparison of the two methods shows interesting results. An orange colour with R=200, G=180 and B=160 gives a hue of 0.0833 (30°) by both methods. However, the Saturation values are completely different at 0.25 and 0.11, and similarly the Value and Intensity values are 0.784 (200/255) and 0.701 (180/255), because of the different definitions.

When hue is calculated close to the sextant boundary the values still remain similar, especially using 8-bit colour depth. For example R=200, G=195, B=160, Hue = 0.146 (HSV) and 0.148 (HSI) i.e. 2 parts in a thousand difference compared with the 4 parts in a thousand difference for 8-bit colour depth.

C Variation of Hue of the same subject with different lighting

Four images (768x512), of the same scene, under different lighting conditions, (Jackson et al, 1994).

Figure A2.3 The same scene illuminated by four different illuminants (Tungsten, White Fluorescent, D50 and D65).

Spot readings were taken, of window size 7x7, of the girl doll's left arm shown in Figure A2.4 as red. The site was chosen, as it was most likely to be a close representation to human skin-colour.

Tables were produced detailing the results of the RGB mean and standard deviation values and HSV mean and standard deviation values for the 49 pixels of the 7x7 window.

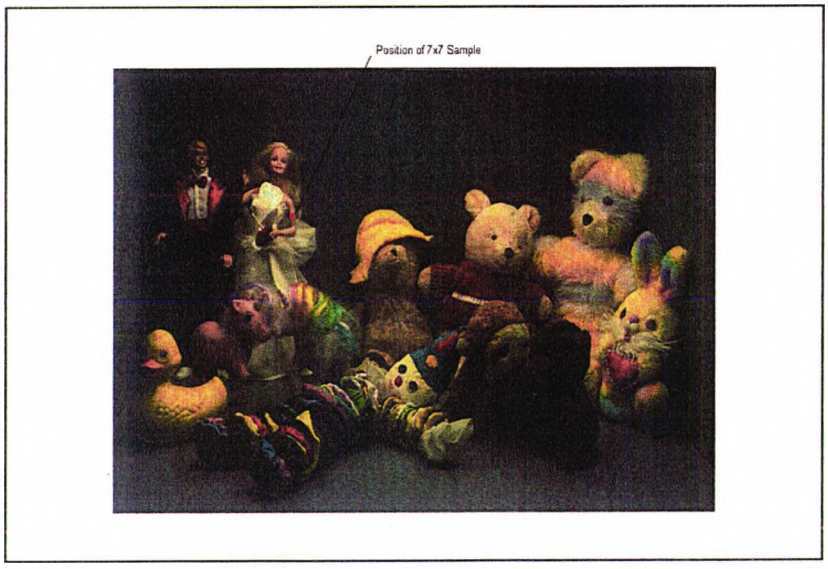


Figure A2.4 Position of the 7x7 sample on the doll's arm shown in red.

	Mean	Standard Deviation
R	201	18.2
G	85	4.8
B	7	2.8
H	0.067	0.004
S	0.964	0.016
V	0.789	0.071

Table A2.1 RGB and HSV values in Tungsten Light

	Mean	Standard Deviation
R	134	26.8
G	25	24.5
B	33	6.1
H	0.168	0.005
S	0.752	0.032
V	0.530	0.100

Table A2.2 RGB and HSV values in White Fluorescent Light

	Mean	Standard Deviation
R	123	19.0
G	104	13.4
B	53	9.2
H	0.123	0.008
S	0.566	0.054
V	0.483	0.074

Table A2.3 RGB and HSV values in Graphics Art D50 Light

	Mean	Standard Deviation
R	146	19.2
G	100	11.0
B	61	8.1
H	0.077	0.006
S	0.584	0.016
V	0.572	0.075

Table A2.4 RGB and HSV values in Textile Dye D65 Light

D Variations in the Hue of Skin-Colour

Samples of skin colour are taken for the right hand, left hand and forehead for five different sequences recorded in a range of environmental conditions.

1 Avatar – a Synthesised Image Sequence

Image sequence, 'Take Mug' supplied by Kennaway, R. of UEA.

The red spots in the top left image of figure A2.5 shows were the samples were taken. Very little difference between the three samples was seen, as would be expected by the rendering technique. The Hue averages at 0.0320, 0.0328 and 0.0341; only a slight variation between the three values. The Hue range of plus/minus two standard deviation from the mean gave good segmentation of the skin-coloured object as seen in figure A2.6. For the optimum range for segmentation the lowest value and the highest value of Hue from the three samples were chosen. Figure A2.6 show the masks resulting from the two segmentation conditions of Hue (left image) and Hue and Saturation (right image).

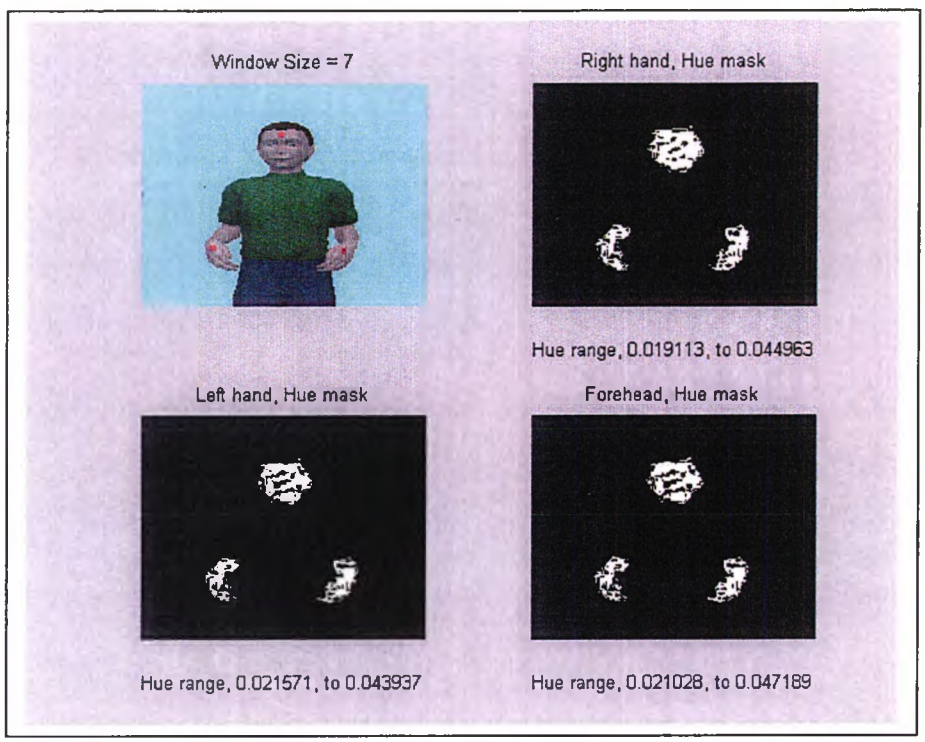


Figure A2.5 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in the 'Take Mug' sequence for a plus/minus two standard deviation from the mean, based on the three samples.

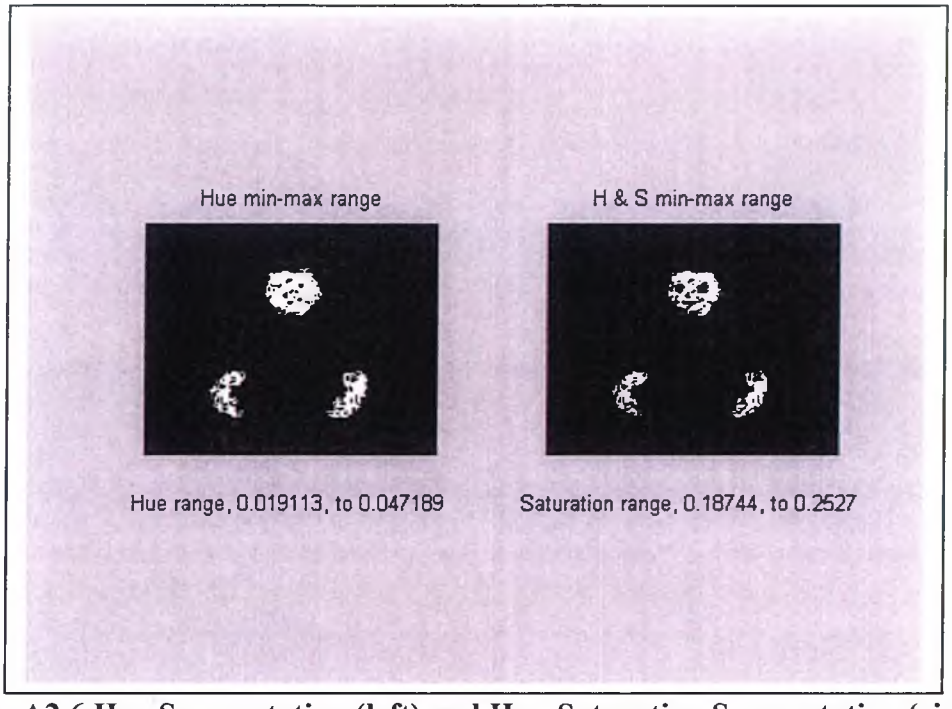


Figure A2.6 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the 'Take Mug' sequence.

2 Complex Scene with Even Illumination

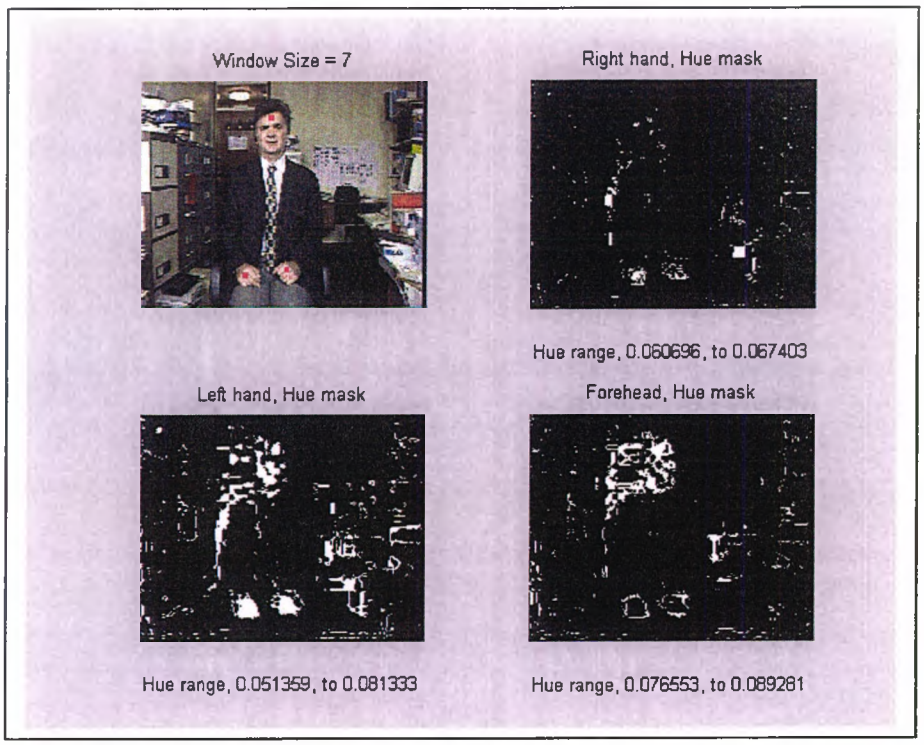


Figure A2.7 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in a complex scene with even illumination for a plus/minus two standard deviation from the mean, based on the three samples.

In this real scene, the variation of the Hue range for the three objects become more apparent, as shown in figure A2.7. The Hues are just a little higher than that of the avatar, and average at 0.0640, 0.0663 and 0.0829. Note that the forehead has a noticeably higher Hue value than the hands. It is noticeable that the wooden door has a very similar Hue range to that of skin and show up in the Hue mask. The Hue & Saturation mask eliminates the door silhouette due to its saturation level being different to skin, as shown in figure A2.8.

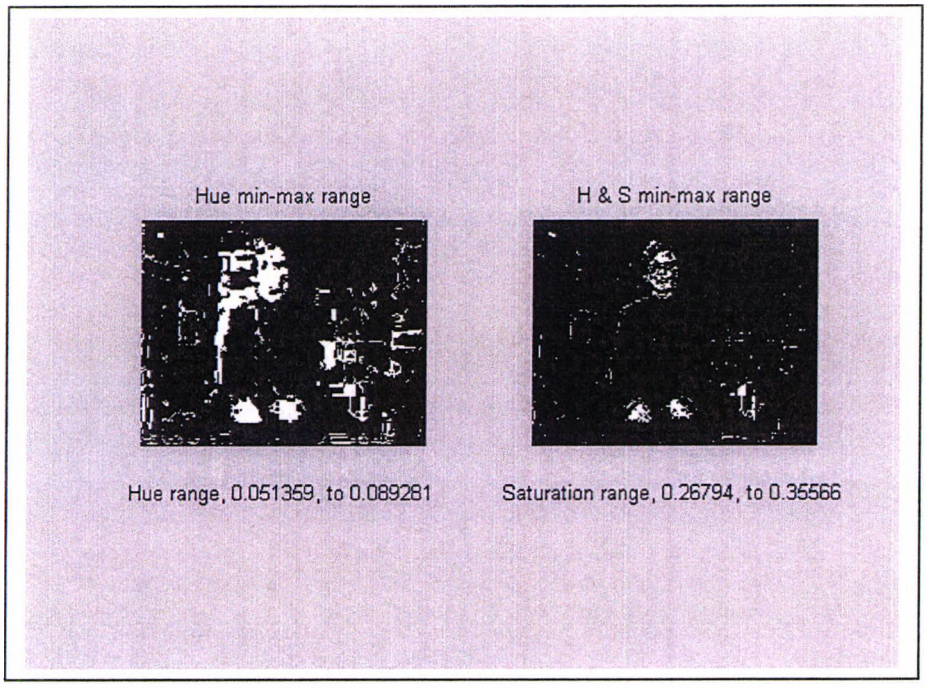


Figure A2.8 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the Complex Scene with Even Illumination.

3 Scene with Low Illumination and Poor White Balance

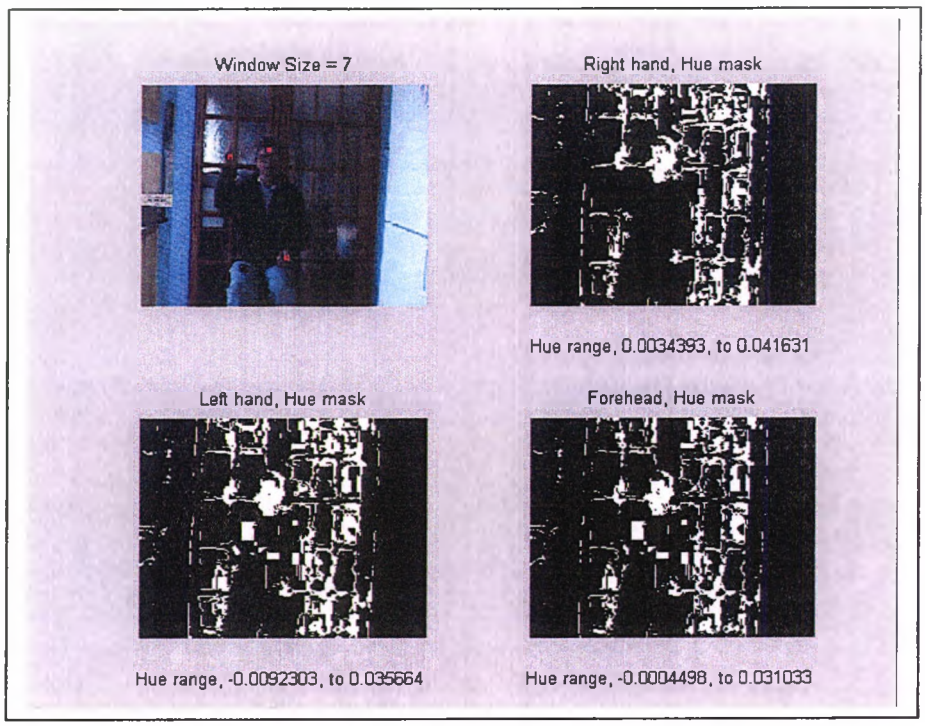


Figure A2.9 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in a Scene with Low Illumination and Poor White Balance for a plus/minus two standard deviation from the mean based on the three samples.

There is much more variability in the hue range in the scene shown in figure A2.9. The right hand has a Hue range very similar to the two previous examples. However, the left hand and forehead have negative Hue values. The negative value is a result of a larger blue component than green in the image due to the white balance mechanism not being adjusted appropriately prior to recording. Some of the white rectangular blocks in the masks are due to the RGB to HSV conversion algorithm response to 'black' and the quantisation due to 'jpg' file format.

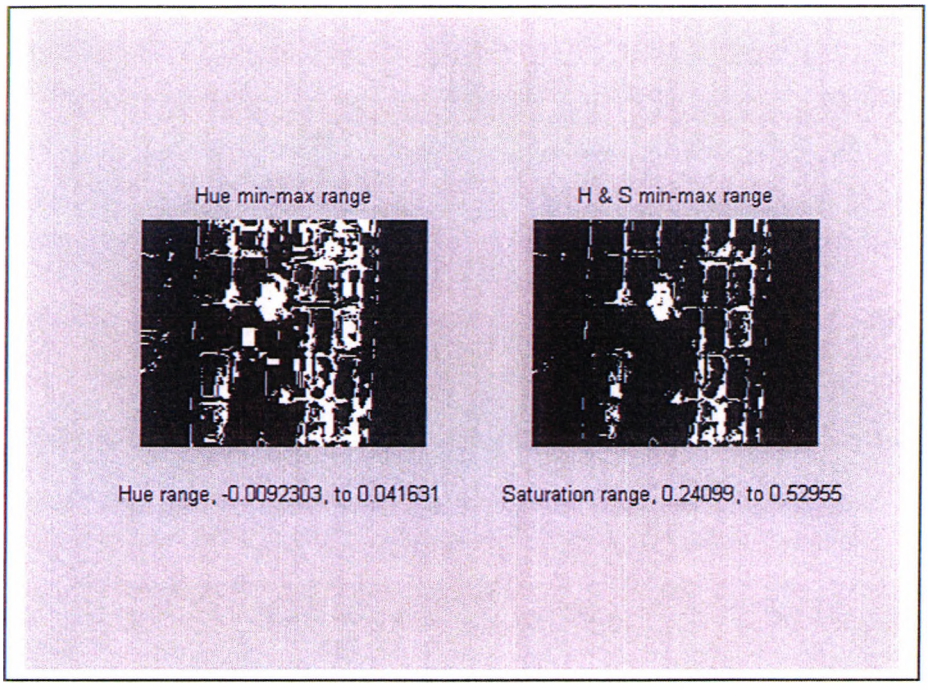


Figure A2.10 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the Scene with Low Illumination and Poor White Balance.

It is noticeable that the wooden door framing is picked up again in the Hue mask and to a lesser extent in the Hue and Saturation mask, as shown in figure A2.10.

4 Challenging Environment and Poor White Balance

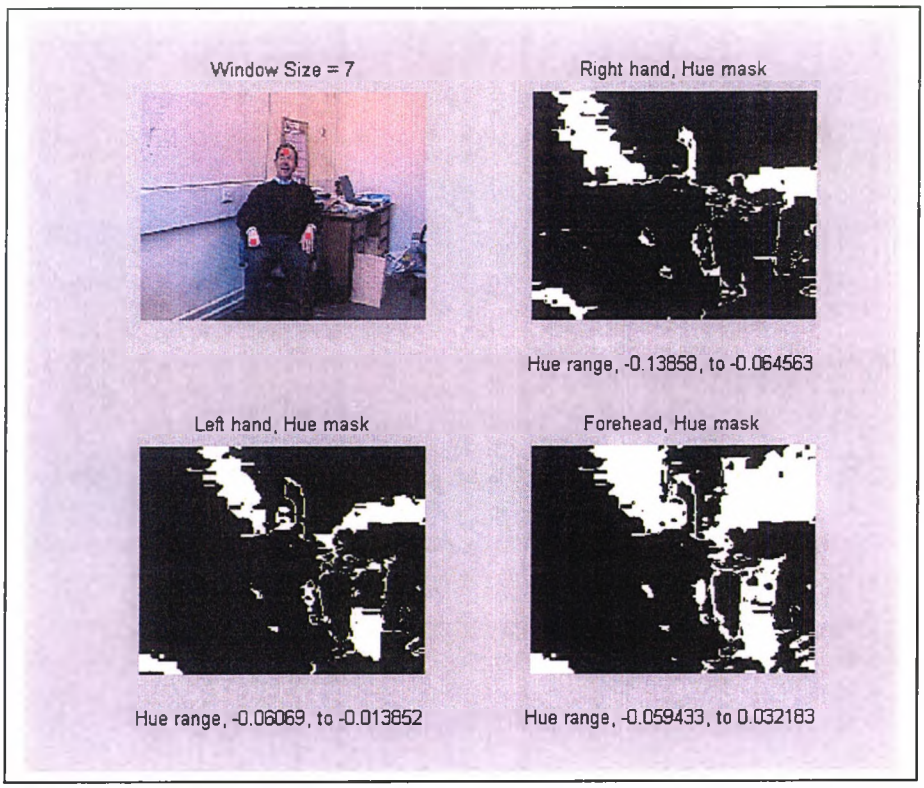


Figure A2.11 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in a Challenging Environment and Poor White Balance for a plus/minus two standard deviation from the mean based on the three samples.

The late afternoon light on the mainly magnolia coloured walls has resulted in the image becoming very pink, as shown in figure A2.11. The majority of the Hue values of the skin samples are now negative, with mean values of -0.1016, -0.0373 and -0.0136. It is noticeable that for the last example that the skin regions of the head and the walls merge together. The Hue-Saturation mask helps segment the head from the wall as shown in figure A2.12.

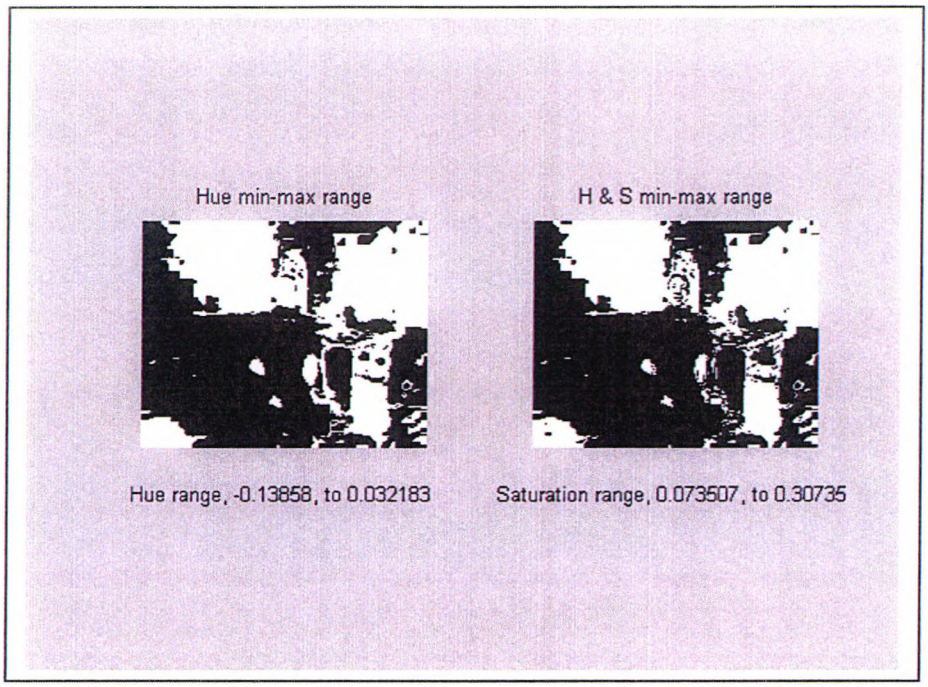


Figure A2.12 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the Scene with a Challenging Environment and Poor White Balance.

5 Publicly available PETS Images

Publicly available sequences, ftp://pets.rdg.ac.uk/PETS-ICVS

Figure A2.13 is an image taken from: - '\data\ScenarioA1\Cam1\image16511.jpg'

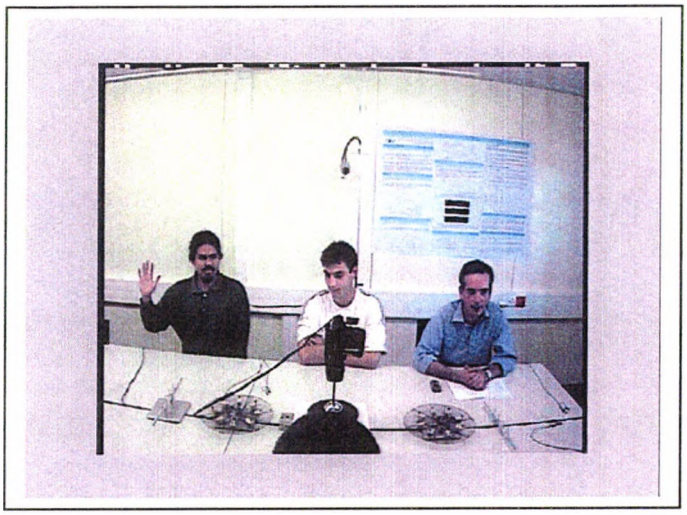


Figure A2.13 Image '\data\ScenarioA1\Cam1\image16511.jpg' from the PETS database.

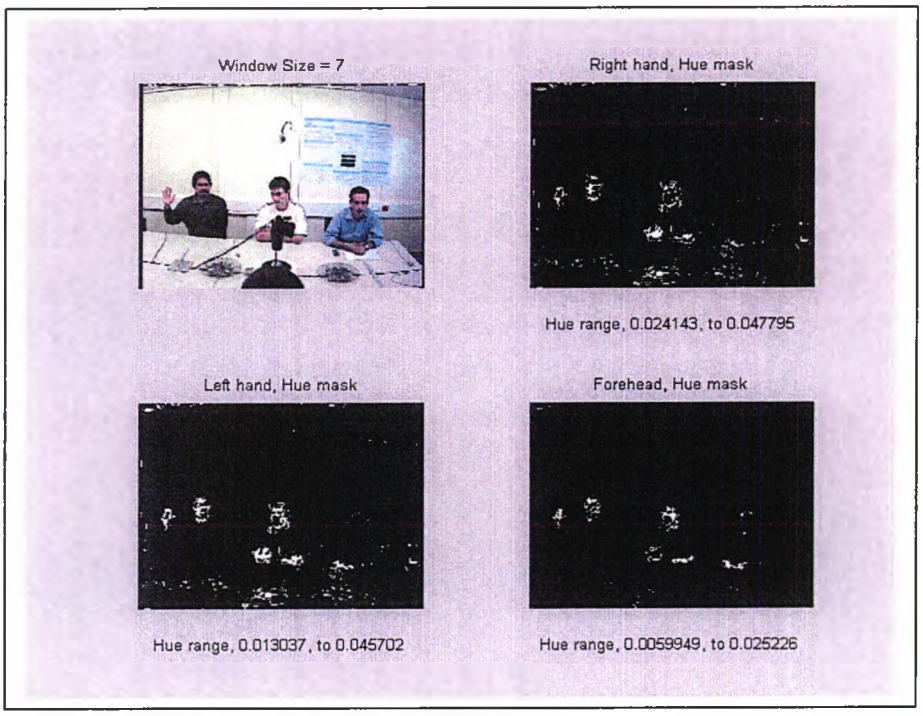


Figure A2.14 Skin-Colour sample positions for the right and left hand and the forehead, with segmentation images of the Hue for a frame in the PETS sequence for a plus/minus two standard deviation from the mean based on the three samples.

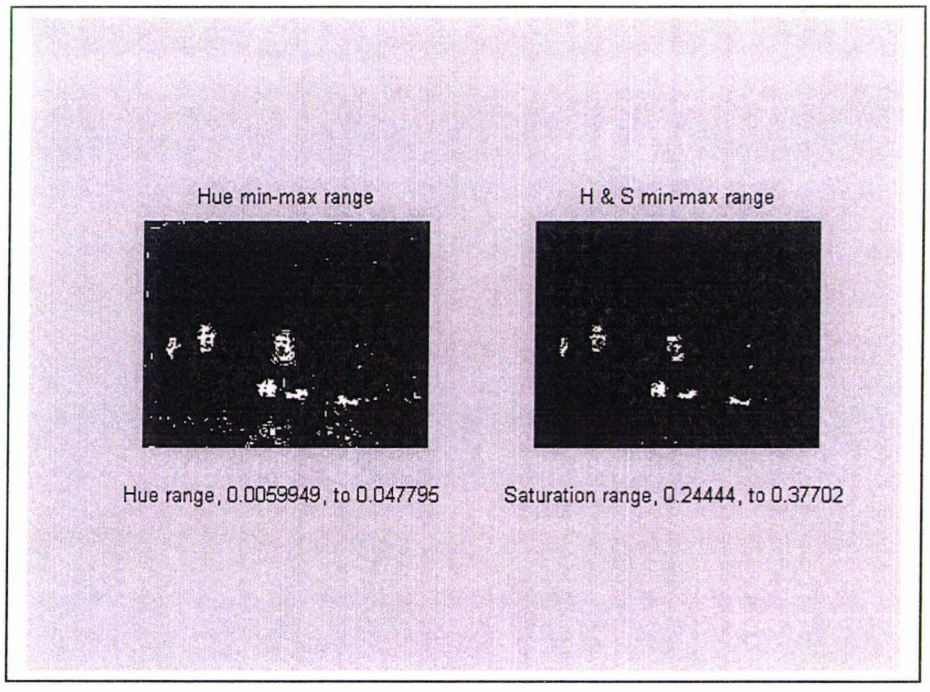


Figure A2.15 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the PETS sequence.

Note that taking 3 sample positions (7x7 window) on the middle person of figure A2.14 segments the skin-coloured regions in the Hue range of 0.006 to 0.048. However, the forehead of the person on the right is not segmented as can be clearly seen in figure A2.15. This is because the hue range extends into the negative region where R>B>G.

In order to segment all skin-coloured regions from the three people the Hue range is extended from –0.048 to 0.033. In this particular example, inclusion of the Saturation range is particularly effective at producing a mask of the skin-coloured regions of just the three people as can be seen in Figure A2.16.

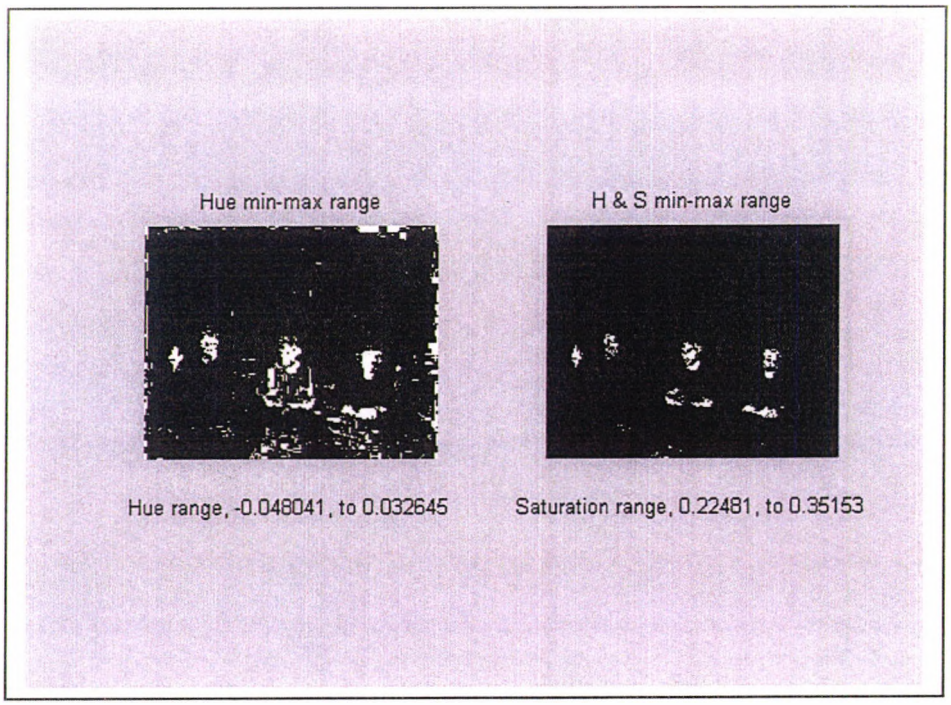


Figure A2.16 Hue Segmentation (left) and Hue-Saturation Segmentation (right) of an image in the PETS sequence to include all skin-coloured regions.

Appendix III – Sequence Parameters

A The Sequence Parameters

Determination of SCM objects can be affected by the skin-colour mask being obtained from a range of values from Hue or Hue-Saturation and whether holes in the masks are morphologically 'filled', or noise objects removed by morphological 'opening'. The eight possible combinations of these parameters are shown in Table A3.1.

Experiment	H or HS	Hole Fill	Opening
E1	H	No	No
E2	H	Yes	No
E3	H	No	Yes
E4	H	Yes	Yes
E5	HS	No	No
E6	HS	Yes	No
E7	HS	No	Yes
E8	HS	Yes	Yes

Table A3.1 Experimental Combinations of H (Hue), HS (Hue-Saturation), Hole 'Fill' and 'Opening'

1 Experiment with Sequence with Low Illumination and Poor White Balance

The sequence consists of some 250 frames and the subject's movements are: -

- Right hand high and then descends to right knee
- Left hand up and down
- Both hands up and down
- Both hands up and across and back
- Pick up remote control of video camera at end.
-

The Hue range was set at –0.005 to 0.043 and the default saturation range was set at 0.01 to 0.95. When combining Hue and Saturation masks the saturation range was set at 0.28 to 0.5. When Prewitt edge templates where used the threshold was set at 0.34. Typical results of the location of the first three SCM objects (red, green and blue crosses) are shown in figure A3.1 and figure 3.2, frames 3 to 19 of the right hand descending.

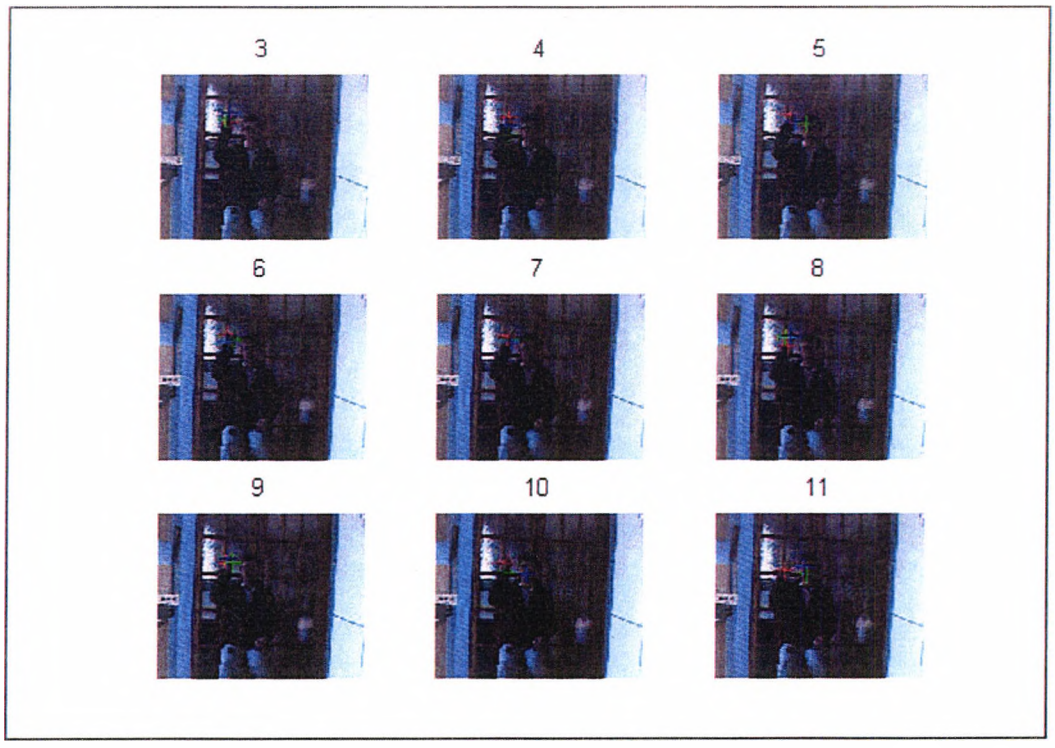


Figure A3.1 Images 3 to 11 of the Low Illumination and Poor White Balance sequence showing the position of the first three SCM objects (red=1st, green=2nd, blue=3rd)

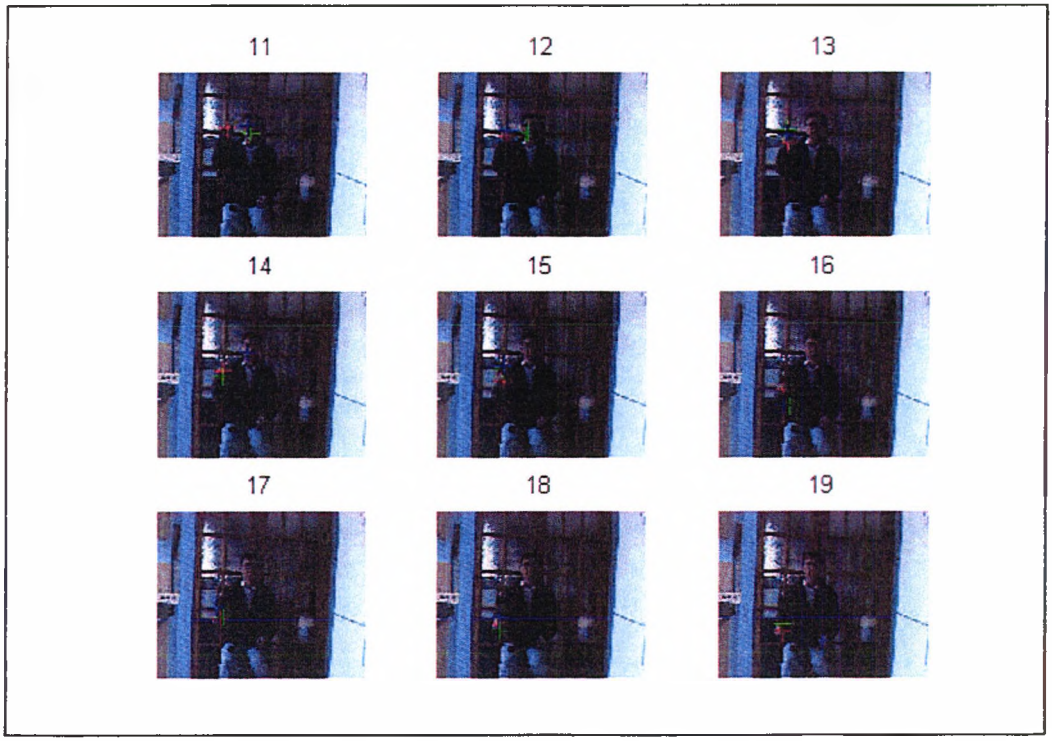


Figure A3.2 Images 12 to 19 of the Low Illumination and Poor White Balance Sequence showing the position of the first three SCM objects (red=1st, green=2nd, blue=3rd)

The difference when using the Hue mask or the Hue-Saturation mask is shown in Figures A3.3 and A3.4. The top right image in each figure shows a red region that shows the skin-colour region defined by Hue or Hue-Saturation. A cyan region is due to motion and the small black regions are due to the logical ANDing of the skin-colour mask and the motion mask. These regions are shown again on the bottom left hand images as 'AND Obj' (SCM). The most noticeable difference between the two figures is the 'Link Obj' (SCMI) regions. Using the Hue-Saturation mask has reduced the size to a more sensible region of the face because of the better segmentation.

The location of the three most significant objects for the SCM and SCMI objects are shown in 2D and 3D space in Figures A3.5 and A3.6, when using the Hue mask, hole-filling and morphological opening conditions specified in Experiment 1 of Table A3.1, respectively for the whole sequence. The following two figures, Figures A3.7 and A3.8 show the result of including the template mask with the SCM and SCMI objects respectively.

In all four figures the gesture space is shown by the location of the objects. Because of the relatively poor lighting conditions and poor skin-colour segmentation the number of objects becomes progressively sparse. A more useful comparison of the performance of the various conditions follows by comparing the allocation of the most significant object to tracking the right and left hand during frames 75 to 120.

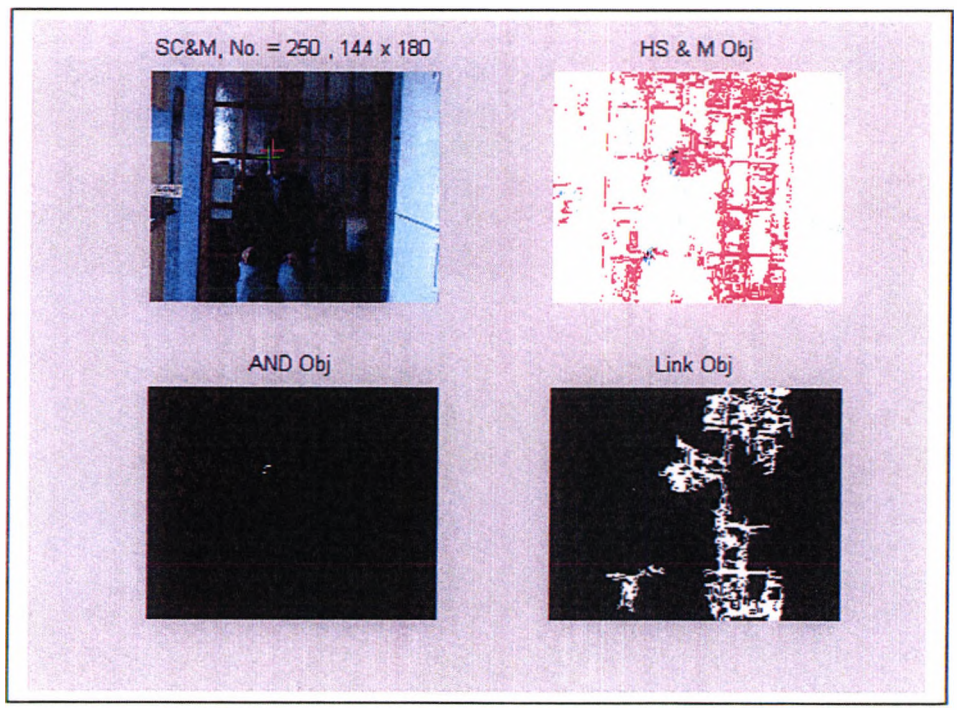


Figure A3.3 Frame 250 showing the centre of gravity of the three SCM objects (red, green and blue crosses) on the head using just the Hue for skin-colour segmentation.

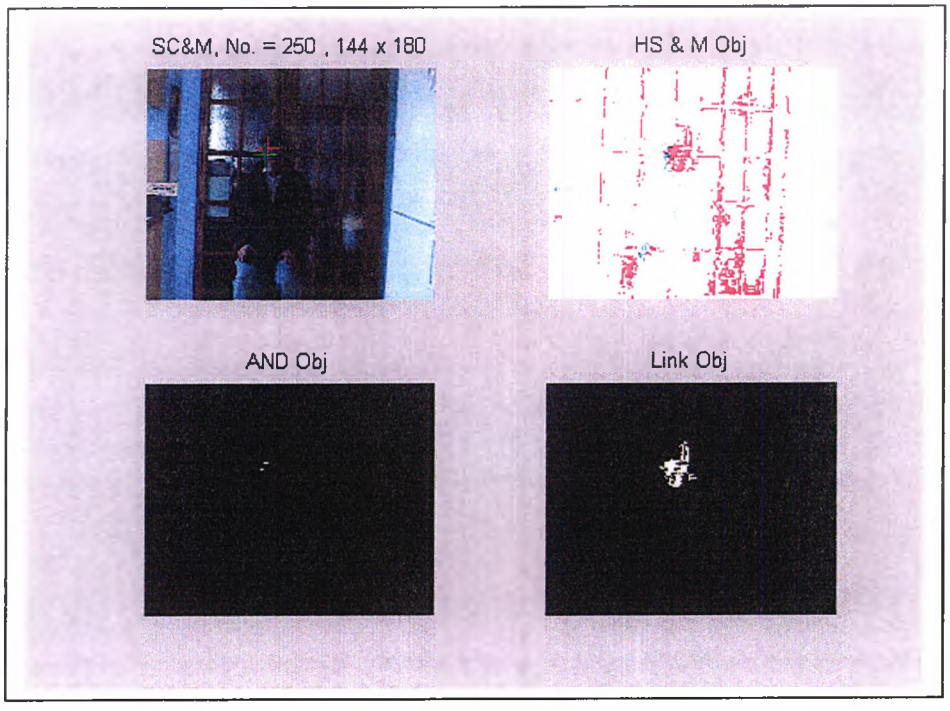


Figure A3.4 Frame 250 showing the centre of gravity of the three SCM objects (red, green and blue crosses) on the head using Hue-Saturation for skin-colour segmentation.

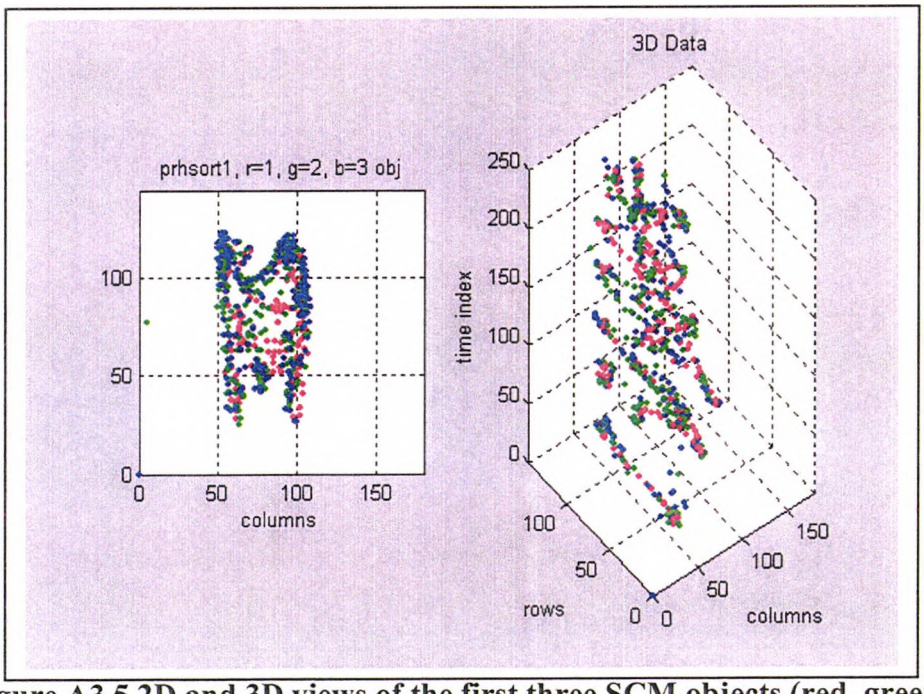


Figure A3.5 2D and 3D views of the first three SCM objects (red, green and blue, respectively), for experiment E1.

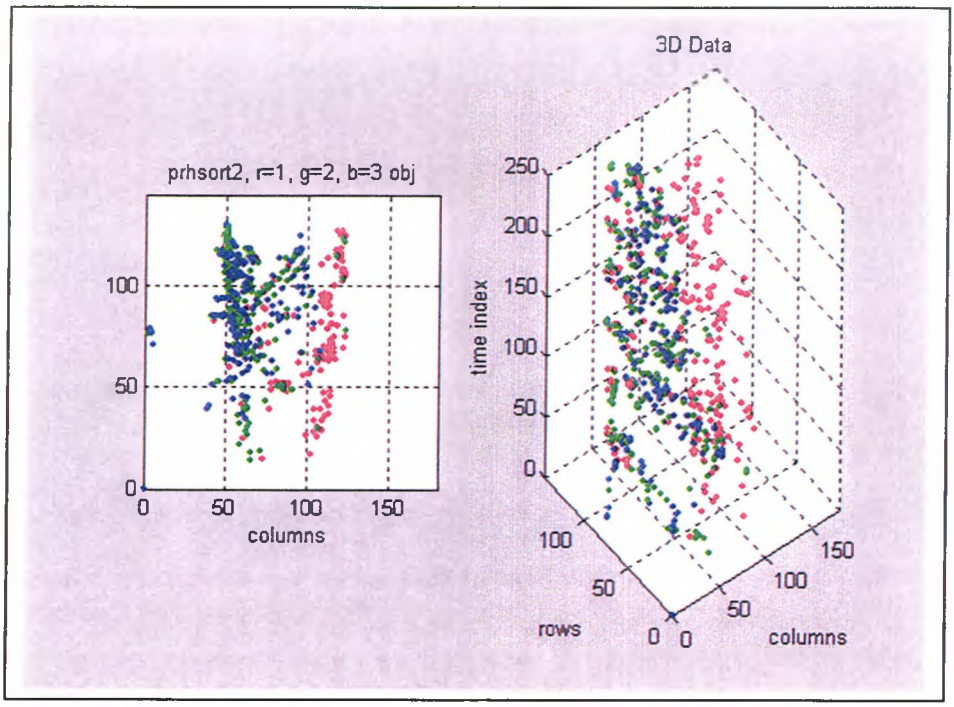


Figure A3.6 2D and 3D views of the first three SCMI objects (red, green and blue respectively), for experiment E1.

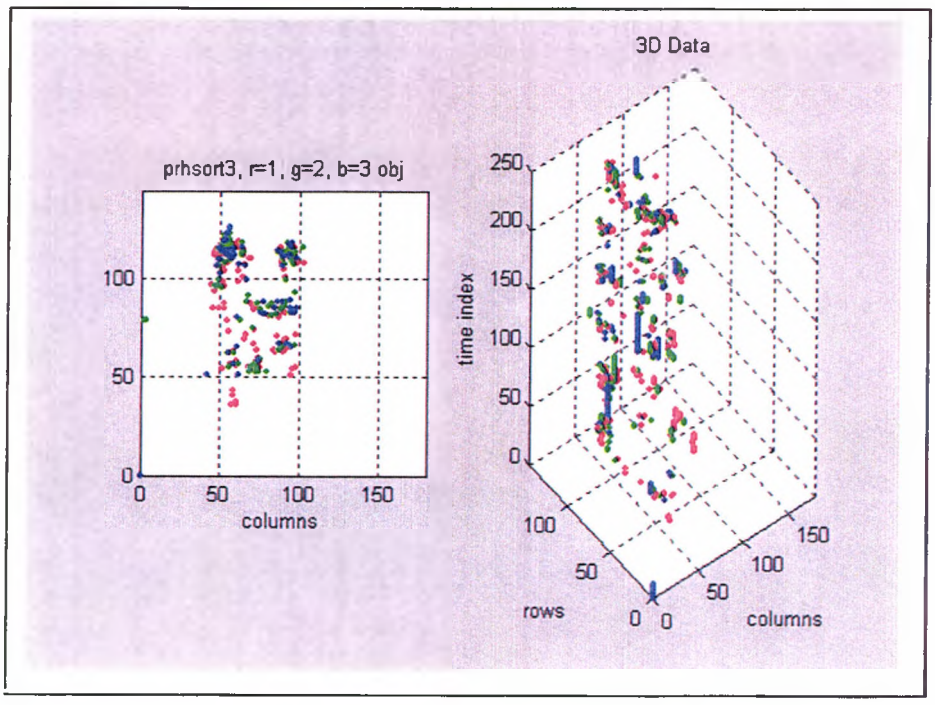


Figure A3.7 2D and 3D views of the first three SCME objects (red, green and blue, respectively), for experiment E1.

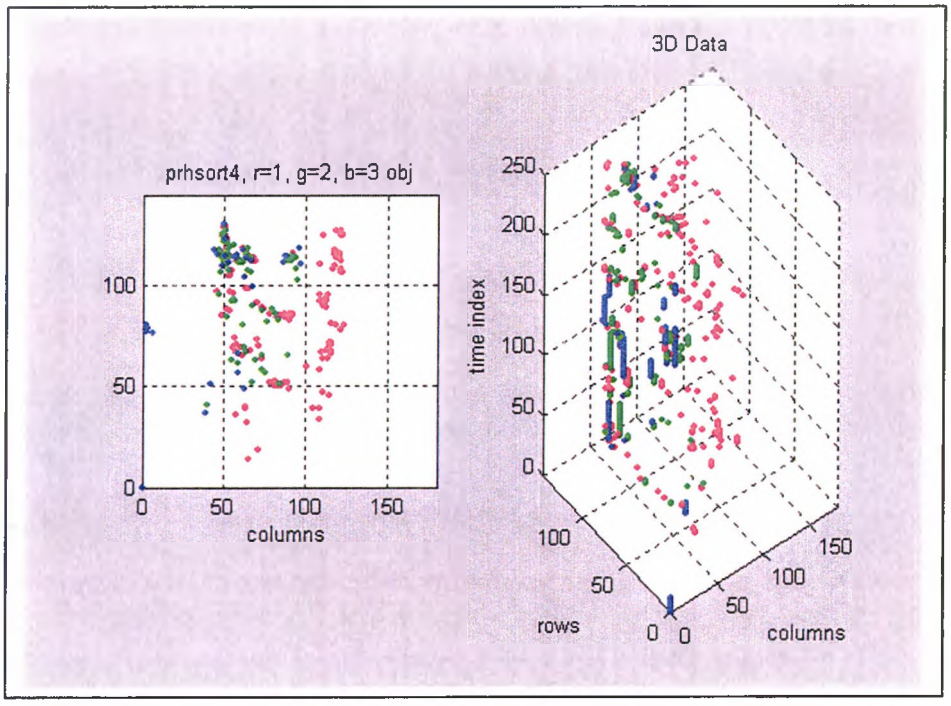


Figure A3.8 2D and 3D views of the first three SCMEI objects from Hue mask (red, green and blue, respectively), for experiment E1.

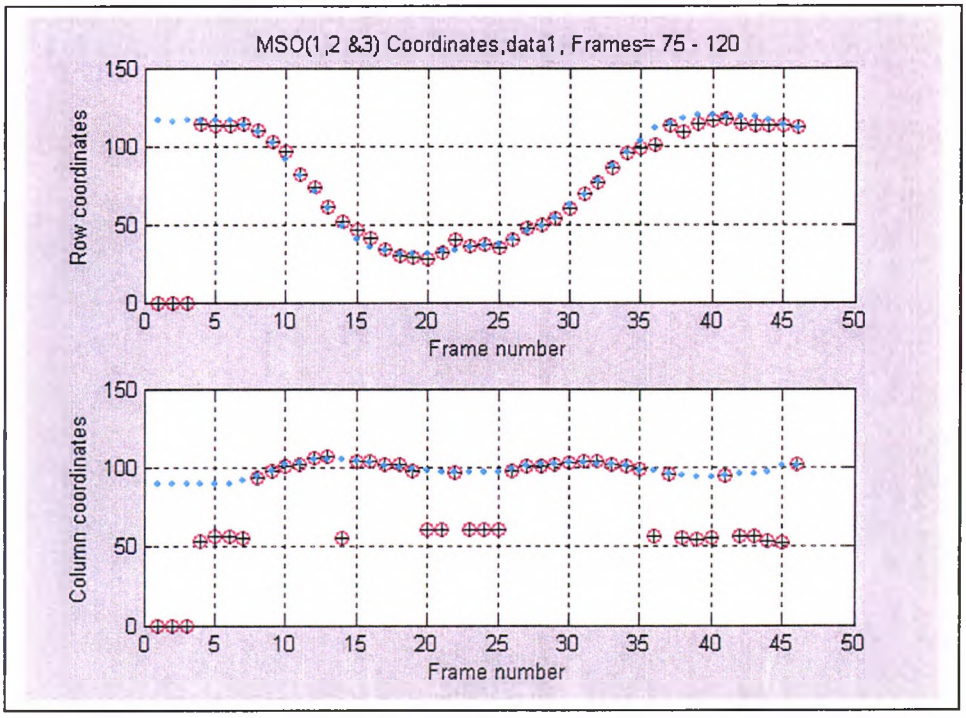


Figure A3.9 Experimental conditions E1 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position)

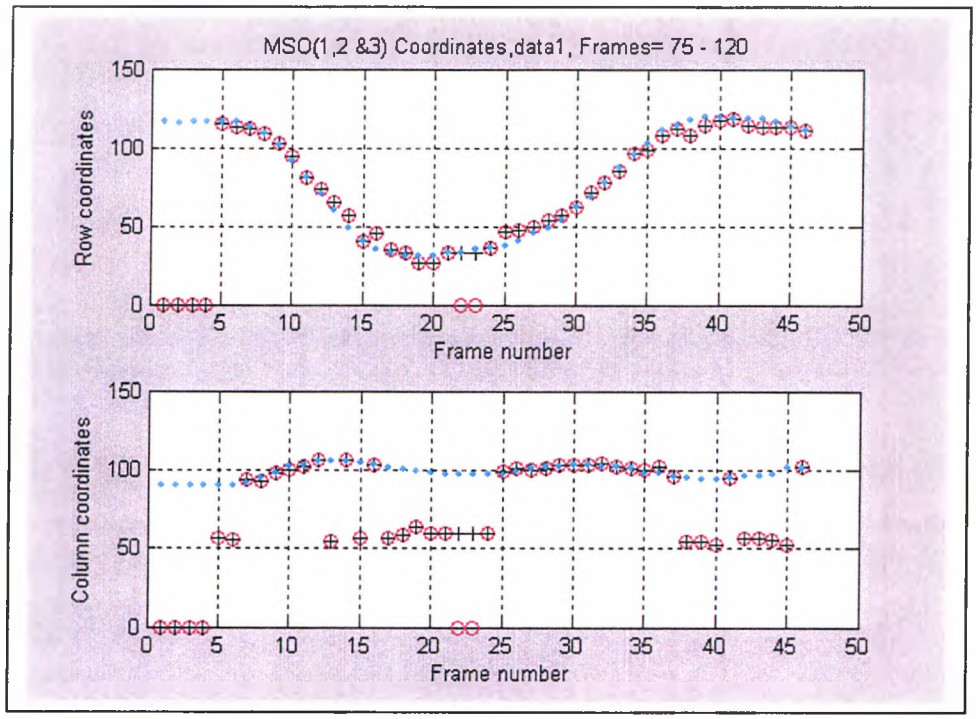


Figure A3.10 Experimental conditions E4 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position)

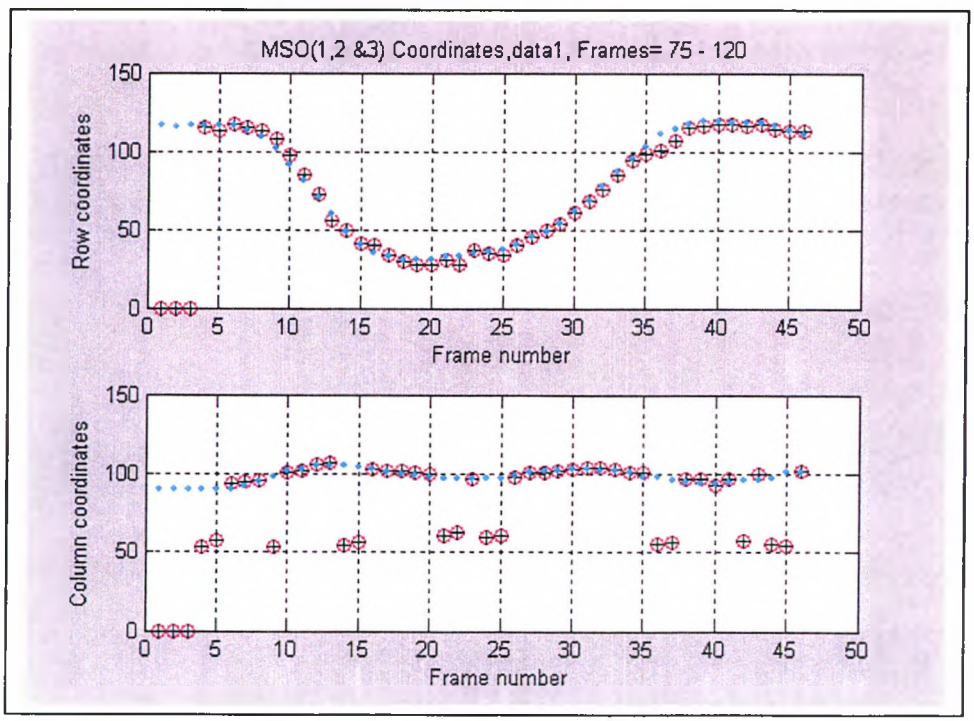


Figure A3.11 Experimental conditions E5 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position)

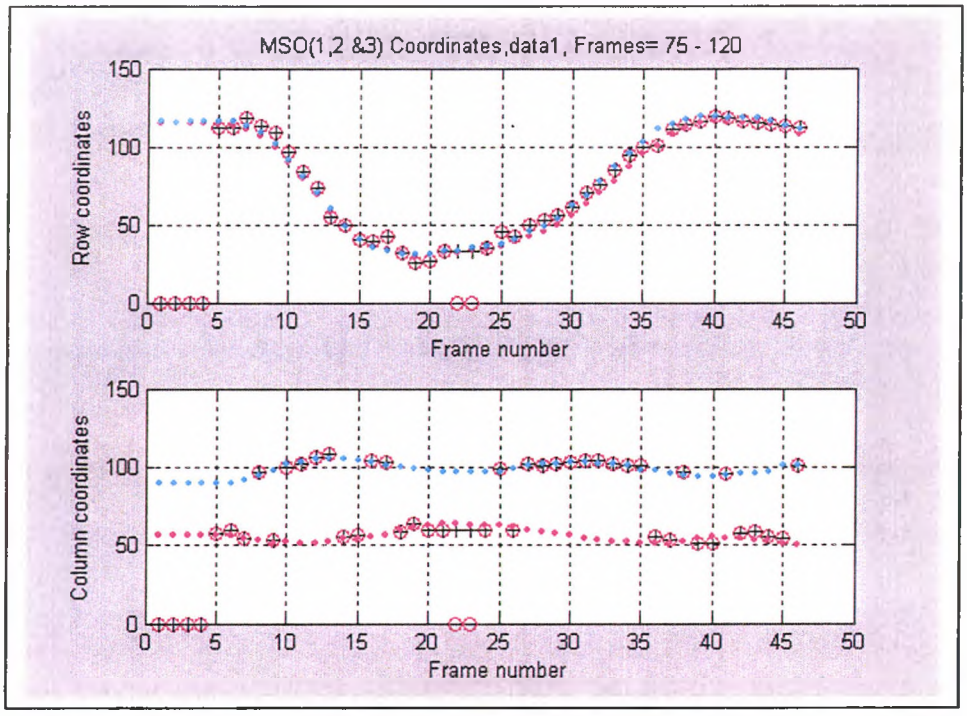


Figure A3.12 Experimental conditions E8 for SCM data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand

There was very little difference in the row coordinates values for the above four sets of experimental results so only column values are used in Table A3.2

Most significant 'scm' object	E1-column	E4-column	E5-column	E8-column
Allocated to Right Hand	25	23	29	20
Allocated to Left Hand	18	17	14	20
Previous values	0	3	0	3

Table A3.2 Comparison of the performance of experiments 1, 4, 5 and 8 for positioning of the 1st 'SCM' object

The distribution of the most significant object is roughly evenly distributed between the right and left hand for all experiments, although there seems a stronger bias for the right hand for experiment E5. Possibly the most significant result from this comparison is that for conditions E4 and E8 when morphological techniques are applied to 'clean' the image, that objects are not generated when the hands are near stationary (frame 4, frame 22 and frame 23).

As a result of poor segmentation of the skin-coloured region the coordinate data for the SCMI objects under experimental conditions E1 (Table A3.1) do not follow the

visually recorded positions of the hands as previously indicated for the SCM most significant object, as shown in Figure A3.13.

In the poor lighting conditions the use of the edge templates with the SCM and SCMI objects produce worse results and there are many instances of objects not being generated for a number of frames. The situation could possibly be improved by constant adjustment of edge template thresholds. However, the overwhelming conclusion of this set of experiments is that the ANDing process is most reliable and performs in poor lighting conditions and it is advisable not to use any morphological processes as important location information can be missed at near stationary conditions. There is only a marginal improvement in object location using the Hue-Saturation mask instead of just the Hue mask on its own. Figure A3.14 shows how the right and left hand can be tracked simultaneously using conditions E5 and the first and second most significant SCM objects, as the visually obtained cyan and magenta tracks show the separate paths of the two hands. The tracking of both hand is signified by the black '+' for the right hand and the blue 'x' for the left hand.

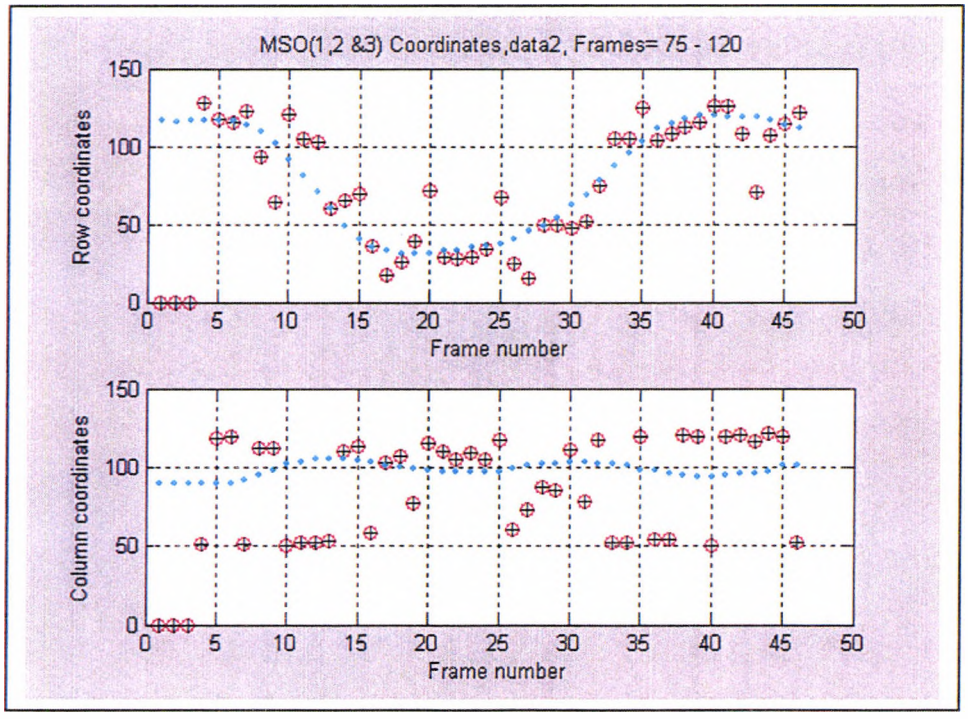


Figure A3.13 Experimental conditions E1 for SCMI data where the most significant object is labelled by a red 'o' and the tracking output signified by a '+' (the cyan dots represent the visually obtained left hand position)

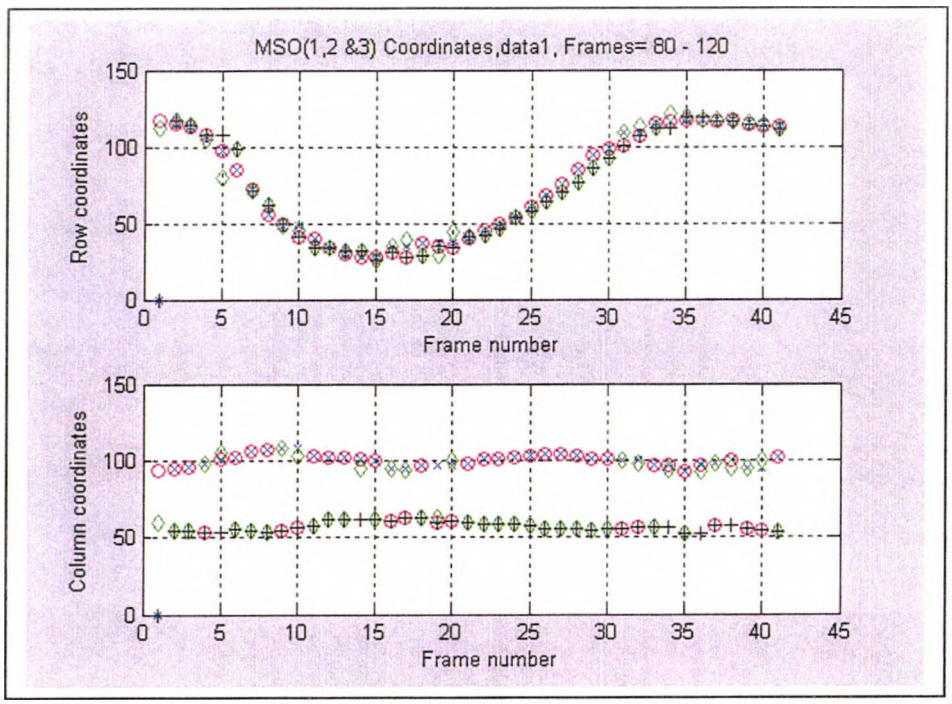
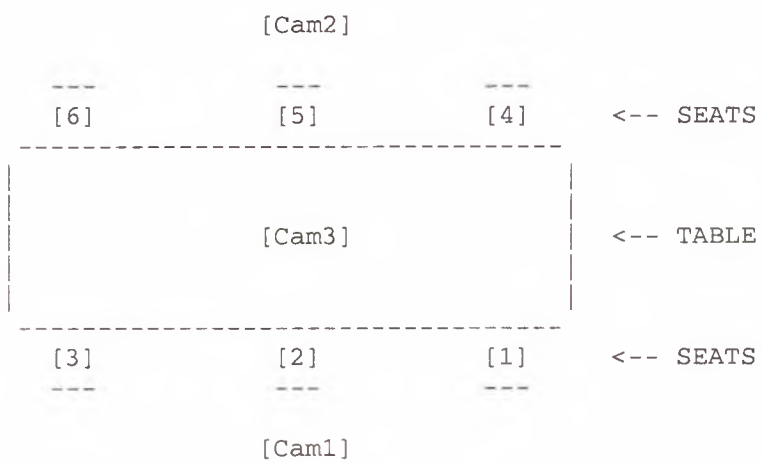


Figure A3.14 Experimental conditions E5 showing 1st (red circle) and 2nd (green diamond) SCM object coordinates. The tracking output is signifies the right hand by the black '+' and the left hand by the blue 'x'. The cyan and magenta dots represent the visually obtained left and right hand coordinate positions, respectively.

2 PETS Sequence

Analysis of a particular sequence – sequence H, frames 16960-17013, with camera recording subjects 6, 5 and 4 as shown in the layout below.



ftp://pets.rdg.ac.uk/PETS-ICVS[Accessed **October 2002**]

X:\images\petsicvs\data\ScenarioA1\Cam1

A montage of images showing the positions of the first three most significant 'scm' objects (red, green and blue crosses) are shown in Figures A3.14 to A3.19.

Figure A3.15 The positioning of the first three most significant SCM objects (red, green and blue, respectively) on frames 16960 to 16968.

Figure A3.16 The positioning of the first three most significant SCM objects (red, green and blue, respectively) on frames 16969 to 16977.

Figure A3.17 The positioning of the first three most significant SCM objects (red, green and blue, respectively) on frames 16978 to 16986.

Figure A3.18 The positioning of the first three most significant SCM objects (red, green and blue, respectively) on frames 16987 to 16995.

Figure A3.19 The positioning of the first three most significant SCM objects (red, green and blue, respectively) on frames 16996 to 17004.

Figure A3.20 The positioning of the first three most significant SCM objects (red, green and blue, respectively) on frames 17005 to 17013.

Table A3.3 classifies the three most significant SCM objects to the positions on the bodies of the three people, labelled 6, 5 and 4 (left to right). The gesturer is number 6 and the position of the object on the right hand, left hand or face is signified by 'rh', 'lh' or 'f'.

Frame No	1st obj = RED '+' on image	2nd obj= GREEN '+' on image	3rd obj = BLUE '+' on image
16960	5	5	6rh
16961	6rh	5	5
16962	5	6rh	6f
16963	6rh	5	6f
16964	6f	5	5
16965	6f	4	6r
16966	6f	4	5
16967	6f	5	4
16968	6rh	4	5
16969	6rh	5	5
16970	6rh	5	5
16971	5	6rh	5
16972	6rh	5	5
16973	5	6rh	5
16974	6rh	5	5
16975	5	6rh	6rh
16976	6rh	5	6rh
16977	5	5	5
16978	5	5	5
16979	5	5	6rh
16980	5	5	6rh
16981	5	6rh	5
16982	5	5	5
16983	5	5	5
16984	5	6rh	5
16985	5	5	5
16986	5	6rh	5
16987	5	6rh	5
16988	5	5	5
16989	5	5	5
16990	5	5	6rh
16991	5	5	6rh
16992	5	5	5
16993	5	5	5
16994	5	5	5
16995	5	5	5
16996	5	5	5

16997	5	6rh	5
16998	6rh	5	5
16999	6rh	5	5
17000	6rh	5	5
17001	5	6rh	5
17002	5	5	6rh
17003	6rh	5	5
17004	6rh	5	5
17005	6lh	6f	6rh
17006	6f	6lh	5
17007	6f	6f	6f
17008	6f	6f	5

Table A3.3 Allocation of the first three SCM objects to subjects 6, 5 and 4 (left to right) and place on body (rh = right hand, lh = left hand and f = face) of the gesturer 6.

Appendix IV - Multirate Ratios

Sample Normalisation Calculation

The selection of the multirate ratios was restricted by the warnings about filter length in the Matlab software. The maximum interpolation value for L was 13 and a similar value, M for decimation. It was found that many ratios could easily be obtained with just one ratio value of L/M. However, better overall results were obtained by the cascading of two ratios i.e. L_1/ M_1 and L_2/ M_2.

Table A4.1 show the ratio values and the resulting error for an optimal length of 64 samples. For example, for a gesture length of 23 the ratio 64/23 = 2.7826. The ratio is calculated from the table as 9/7 x 13/6 = 2.7857, a difference of 0.0031 (0.11%).

Number of samples	L_1	M_1	L_2	M_2.	error	overshoot
10	8	5	4	1	0	0
11	5	1	7	6	0.015	1
12	4	3	4	1	0	0
13	9	5	11	4	0.027	1
14	8	7	4	1	0	0
15	8	5	8	3	0	0
16	4	1	1	1	0	0
17	3	2	5	2	-0.015	1
18	4	3	8	3	0	0
19	13	9	11	4	0.0019	1
20	8	5	2	1	0	0
21	8	7	8	3	0	0
22	7	6	5	2	0.0076	1
23	9	7	13	6	0.0031	1
24	8	3	1	1	0	0
25	8	5	8	5	0	0
26	8	7	13	6	0.0147	1
27	9	7	11	6	-0.0132	1
28	2	1	8	7	0	0
29	11	5	1	1	-0.0069	1
30	4	3	8	5	0	0
31	9	8	11	6	-0.002	1
32	2	1	1	1	0	0
33	7	6	5	3	0.0051	1
34	9	8	5	3	-0.0074	1
35	8	7	8	5	0	0
36	4	3	4	3	0	0
37	13	12	8	5	0/0036	1
38	13	9	7	6	0.001	1
39	13	9	8	7	0.0098	1

40	8	5	1	1	0	0
41	13	12	13	9	0.0038	1
42	8	7	4	3	0	0
43	11	9	11	9	0.0055	1
44	9	8	9	7	-0.0081	1
45	7	6	11	9	0.0037	1
46	13	12	9	7	0.0016	1
47	7	6	7	6	-0.0005	1
48	4	3	1	1	0	0
49	8	7	8	7	0	0
50	9	7	1	1	0.0057	1
51	13	12	7	6	0.009	1
52	13	12	8	7	0.0073	1
53	13	12	9	8	0.0112	1
54	13	12	13	12	-0.0116	0
55	13	12	13	12	0/01	0
56	8	7	1	1	0	0
57	9	8	1	1	0.0022	1
58	13	12	1	1	-0.0201	-1
59	13	12	1	1	-0.0014	0
60	13	12	1	1	0.0167	1
61	13	12	1	1	0.0342	3
62	1	1	1	1	-0.0323	-2
63	1	1	1	1	-0.0159	-1
64	1	1	1	1	0	0
65	1	1	1	1	0.0154	1
66	1	1	1	1	0.0303	2
67	12	13	1	1	-0.034	-2
68	12	13	1	1	-0.02	-1

Table A4.1 Calculation of ratios for normalisation of gesture length to 64.

Appendix V– Interpretation of Harmonic Data

1 DFT Calculation

Calculating the harmonic output from the sequence 1, 0, 0, 1.

Harmonic, k	n = 0	n = 1	n = 2	n = 3	Σ
	1	0	0	1	Harmonic value
k = 0, X(0)	$1.e^{-j(2\pi/4).0.0}=1$	0	0	$1.e^{-j(2\pi/4).3.0}=1$	2
k = 1, X(1)	$1.e^{-j(2\pi/4).0.1}=1$	0	0	$1.e^{-j(2\pi/4).3.1}=+j$	1 + j
k = 2, X(2)	$1.e^{-j(2\pi/4).0.2}=1$	0	0	$1.e^{-j(2\pi/4).3.2}= -1$	0
k = 3, X(3)	$1.e^{-j(2\pi/4).0.3}=1$	0	0	$1.e^{-j(2\pi/4).3.3}=-j$	1 - j

Table A5.1 Calculation of harmonic components for data 1, 0, 0, 1

2 Fourier Descriptor Theory (Source: Lin et al., 1990)

A closed 2-D contour with perimeter η is illustrated in Figure A5.1. A point p starts from some arbitrary location on the perimeter and moves along the contour a distance l to s. The coordinates of p can be obtained by defining a parameter t as $2\pi\ell/\eta$, these two periodic functions can be expressed by Fourier expansions in matrix form as: -

$$\begin{bmatrix} x(t) \\ y(t) \end{bmatrix} = \begin{bmatrix} a_0 \\ d_0 \end{bmatrix} + \sum_{k=1}^{\infty} \begin{bmatrix} a_k & b_k \\ c_k & d_k \end{bmatrix} \begin{bmatrix} \cos kt \\ \sin kt \end{bmatrix}, \qquad (1)$$

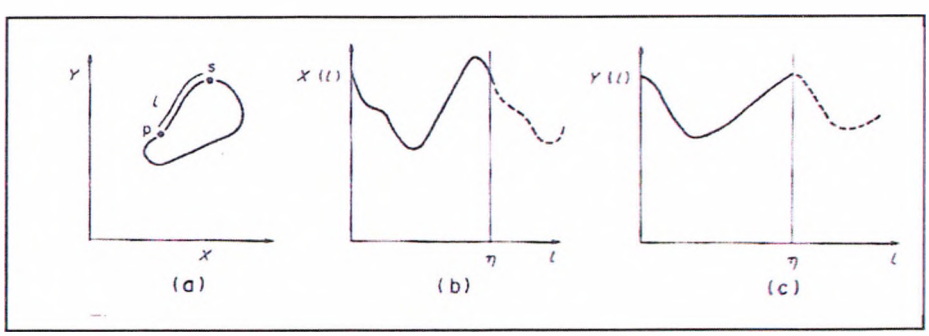


Figure A5.1 (a) A 2-D closed contour. (b, c) Periodic functions $X(l)$ and $Y(l)$ for the contour of (a) (Source: Lin et al.)

Objects can be placed at different locations with arbitrary orientations will be changed as shown in the example in Figure A5.2.

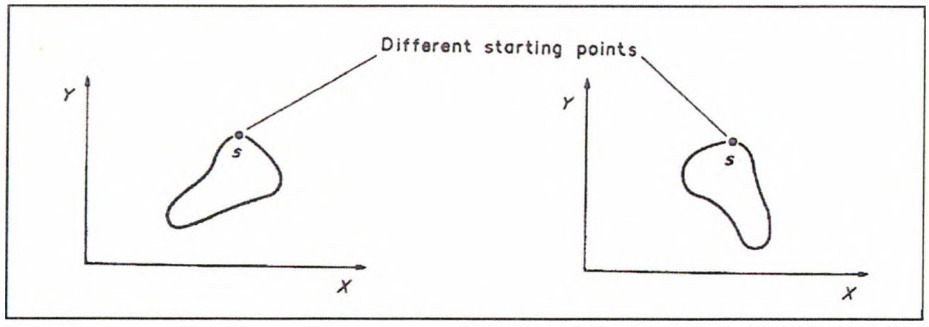


Figure A5.2 Different starting points due to different orientations (Source: Lin et al.)

In equation (1), a_0, d_0 are the mean values of the coordinates $X(l)$ and $Y(l)$, respectively, which indicate the geometric centre of the contour.

The above term for the k^{th} ellipse is: -

$$\begin{bmatrix} x_k(t) \\ y_k(t) \end{bmatrix} = \begin{bmatrix} a_k & 0 \\ 0 & b_k \end{bmatrix} \begin{bmatrix} \cos kt \\ \sin kt \end{bmatrix}$$

This equation of the k^{th} ellipse can be decomposed and related to the orientation and the phase shift of the ellipse. Lin (1990) shows that the ellipses as in Figure A6.3 can be described by the equation: -

$$\begin{bmatrix} x_k(t) \\ y_k(t) \end{bmatrix} = \begin{bmatrix} \cos\theta_k & -\sin\theta_k \\ \sin\theta_k & \cos\theta_k \end{bmatrix} \begin{bmatrix} A_k & 0 \\ 0 & B_k \end{bmatrix} \begin{bmatrix} \cos\varphi_k & -\sin\varphi_k \\ \sin\varphi_k & \cos\varphi_k \end{bmatrix} \begin{bmatrix} \cos kt \\ \sin kt \end{bmatrix}$$

where A_k and B_k represent the major and minor axis of the ellipse, the starting phase of the ellipses is θ_k and the phase is defined as $\varphi_k = 2\pi\delta / perimeter$, being the phase from the major axis to the position corresponding to $t=0$.

An alternative interpretation of the Fourier Descriptors is a set of oriented ellipses as shown in Figure A5.4. A shape can be viewed geometrically as the locus generated by the case of three ellipses. Each ellipse has a fixed orientation, with the centre of the $k^{th}+1$ ellipse, C_k, revolving around the k^{th} ellipse. The revolving frequency of C_k is k times that of C_1. The locus of the last point C_3 forms the contour. It is often found that only a few of the low order harmonics are necessary to synthesise the object and the higher harmonics usually being very high frequencies and low amplitudes do not have a significant contribution to the overall shape of the object.

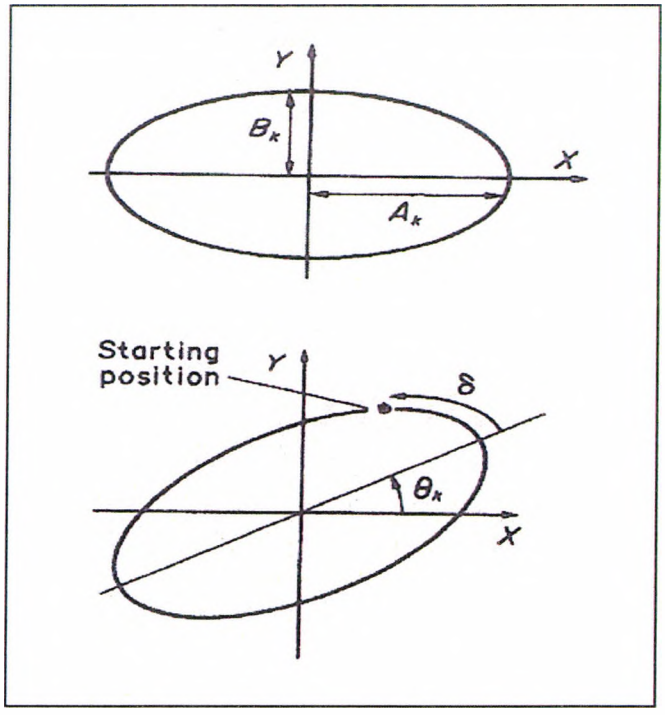


Figure A5.3 The rotation and starting phase of an ellipse. Source (Lin et al.)

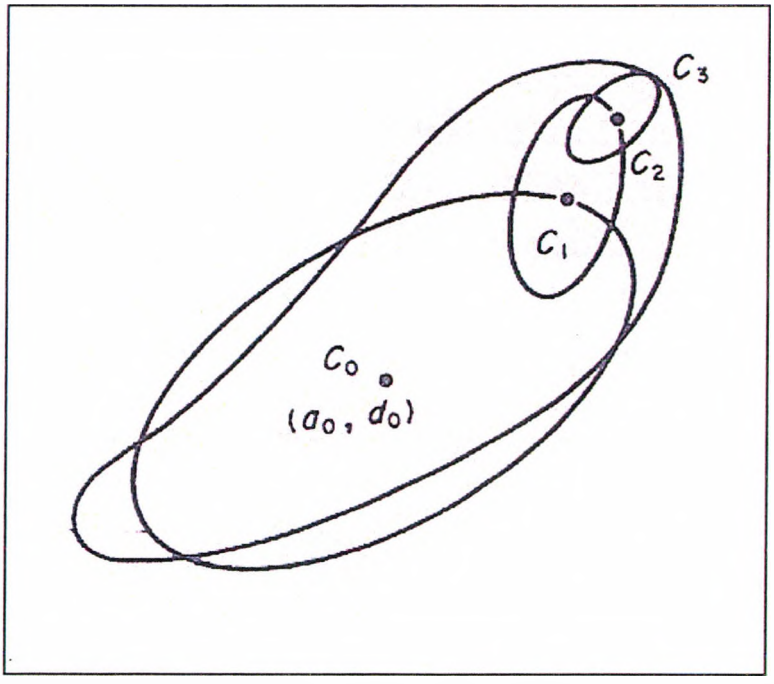


Figure A5.4 Three harmonic elliptic descriptions. Source (Lin et al.)

3 Visualisation of exponential equations in 1D, 2D and 2DT

Two ellipses are added together and result shown in Figure A5.5. The characteristics of the ellipses are defined according to Lin's (1990) work, with the major axis, labelled A and the minor axis labelled B. In this example the first and second harmonics have major and minor axis of A_1=0.5, B_1=0.3; A_2=0.2, B_2=0.1 respectively, with an offset of 1.0 for the second harmonic ellipse. The top left picture shows the spatial domain representation of the two ellipses. The top-right shows the change of 'y' with time, clearly showing the offset bias. The bottom-left picture shows the change of 'x' with time and the bottom right shows a 2DT view.

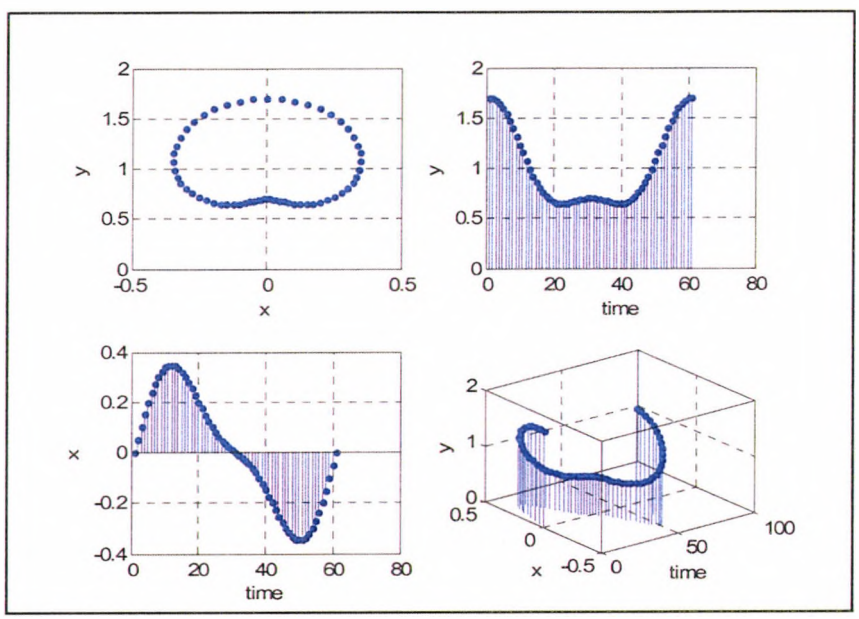


Figure A5.5 Pictures of 4 views of the summation of two ellipses of different frequencies with an offset. The top left picture shows the spatial domain representation; the top-right shows the change of 'y' with time, clearly showing the offset bias; the bottom-left picture shows the change of 'x' with time and the bottom right shows a 2DT view.

The following figure, Figure A5.6 shows that the 2DT image changes with viewing angle and does not capture a constant 2D shape or interpretation.

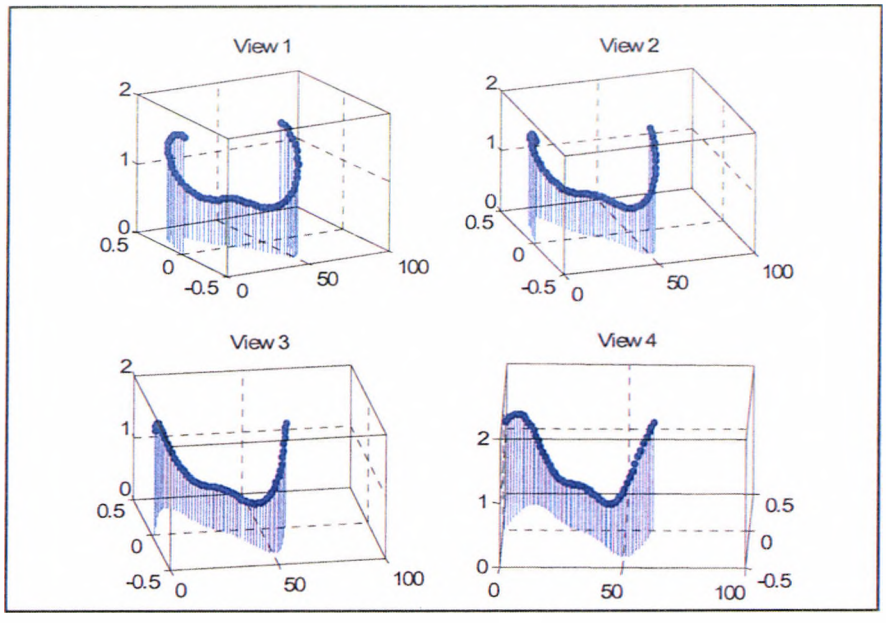


Figure A5.6 Different azimuth but the same elevation of two ellipses with offset (Azimuth -View 1 =30°, View 2 =20°, View 3 =10°, View 4 =0°; Elevation = 30°)

Synthesis of waveforms

The power of Fourier synthesis is often shown by taking just a few low order harmonics and adding them together to show that a good representation of the original waveform can be made. It can be shown that recombining the first three odd harmonics, as the even order harmonics are zero, from square wave analysis shows how close the synthesis is to the original square waveform. For the purposes of this investigation it is useful to visualise the waveform both in the 2D (spatial), 1D and 2DT domains. A sine wave is generated in the 'x' domain by adding a phase shift of minus 90° to the cosine wave of the real part of the exponential, as given previously by Euler's equation. The sine wave is shown in the 'x' vs. time plot (bottom-left picture) of Figure 5.7. There is no component on the 'y' axis (top-right picture of Figure 5.7) and a 2DT view is shown in the bottom-right picture. The top-right picture shows the equal amplitudes of the positive (red) and negative (black) sequence component and the overall appearance based result (blue) which is just a straight line because of the zero 'y' component. The third harmonic, at appropriately lower amplitude is shown in Figure 5.8. The result of adding the first three odd harmonics is shown in Figure 5.9.

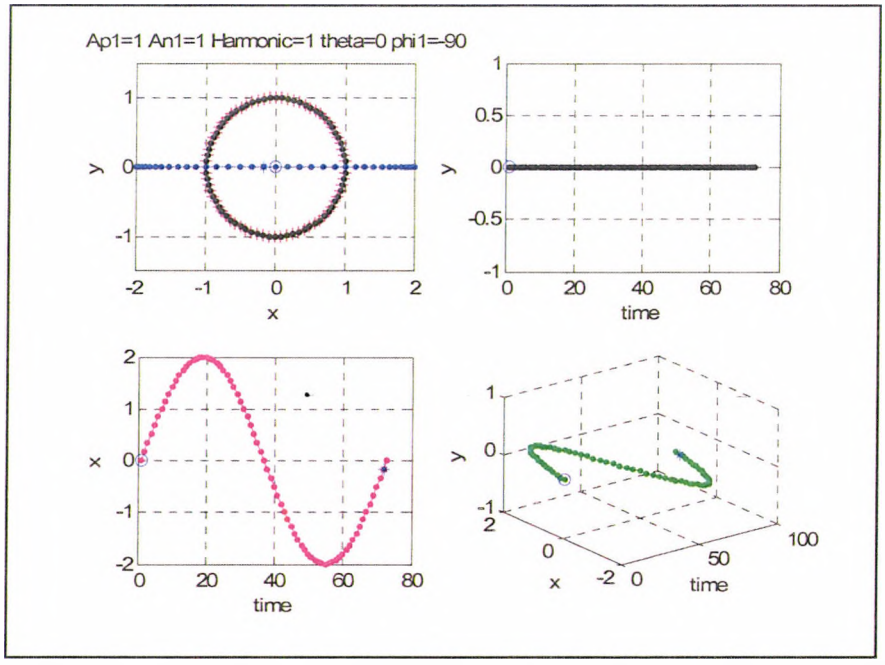


Figure A5.7 Four views (top-left spatial domain, x-y; top-right, 'y' vs. time; bottom-left, 'x' vs. time; bottom-right, 2DT domain) of the first harmonic. The starting point-and end-point of the time sequence indicated as 'o' and '*' (blue) respectively.

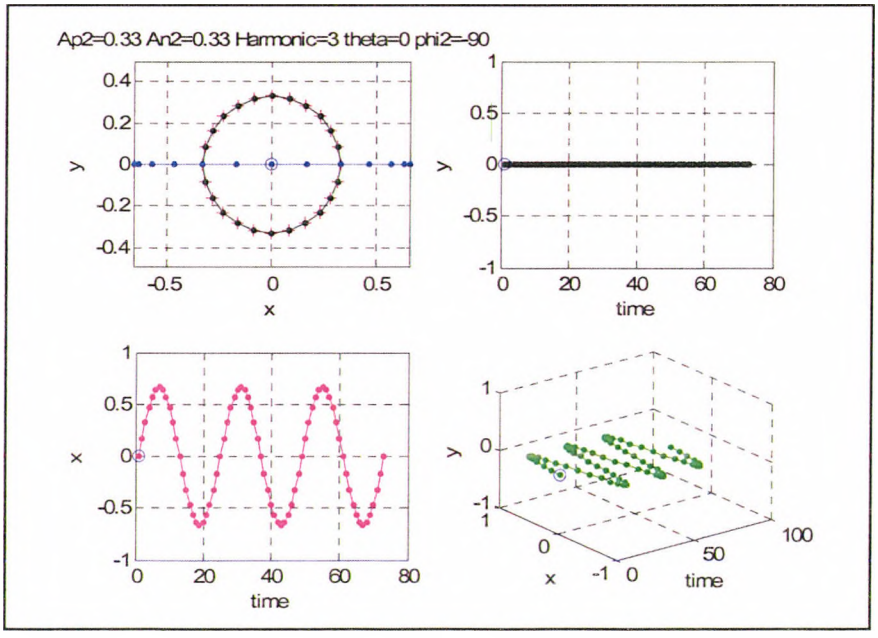


Figure A5.8 Four views (top-left spatial domain, x-y; top-right, 'y' vs. time; bottom-left, 'x' vs. time; bottom-right, 2DT domain) of the third harmonic. Rotating positive sequence (red) and negative (black) sequences and the resulting (blue) ellipse.

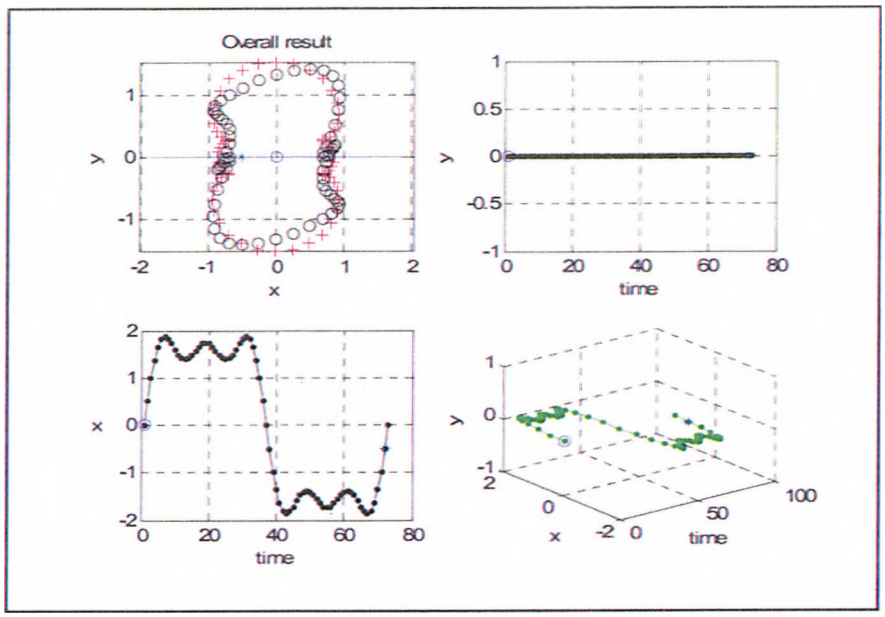


Figure A5.9 Four views of combining the first, third and fifth harmonics. Rotating positive (red) and negative (black) sequences and the resulting (blue) ellipse. The starting point-and end-point of the time sequence indicated on all pictures as 'o' and '*' (blue) respectively.

4 Simulated gesture trajectories

Different possible gesture paths were simulated. Shallow and deep arcs were plotted and called 'concave'. Another trajectory with the arc in the opposite 'sense' to concave was termed a 'convex' trajectory. Additionally, 'figure of eight' trajectory and a trajectory with many oscillations were generated. The coordinates used for these simulated trajectories are shown below.

Data Sets for Trajectories

A Shallow arc or concave trajectory

x=10 23 33 46 58 69 82 96 118 130 144 160 157 150 143 135 125 115 105 92 77 62 45 30

y=200 158 137 118 102 86 72 57 42 30 19 10 30 50 63 78 95 105 122 138 153 166 180 188

B Deeper arc or concave trajectory

x= 32 50 72 87 102 120 135 145 155 160 165 166 166 166 166 165 160 155 145 135 120 102 87 72 50 32

y = 192 184 173 164 152 140 125 110 90 75 52 35 12 12 35 52 75 90 110 125 140 152 164 173 184 192

C Convex trajectory

x= 10 23 33 46 58 69 82 96 118 130 144 160 160 144 130 118 96 82 69 58 46 33 23 10

y = 200 158 137 118 102 86 72 57 42 30 19 10 10 19 30 42 57 72 86 102 118 137 158 200

D Figure of eight trajectory

x= 10 27 45 60 71 80 90 93 96 100 108 123 135 160 162 153 140 128 115 72 53 40 30 20 13 10

y = 200 197 188 176 162 146 130 95 78 61 43 27 15 10 38 58 74 82 88 95 102 113 130 142 162 181

E An Oscillating -Three figures of eight trajectory

x = 10 30 38 42 50 66 82 88 98 102 135 140 160 146 125 110 112 98 77 62 62 52 32 26

y = 200 187 176 154 139 133 119 96 79 70 52 22 10 35 42 62 81 100 105 122 140 159 168 181

Simulated trajectories and Harmonic content

Shallow arc or concave trajectory

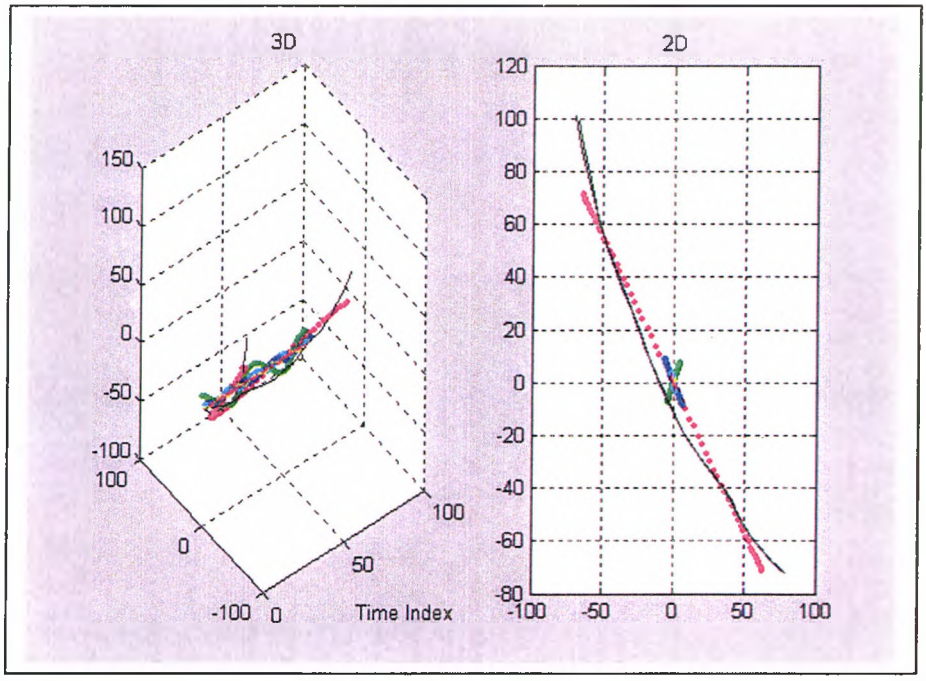


Figure A5.10 2DT and 2D views of the harmonics of a shallow arc or concave Trajectory

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-7.2	1.0	7.2	-131.8
2	0.093	-13.8	0.092	13.8	72.7
3	0.111	-20.7	0.113	20.7	-66.3
4	0.038	-24.7	0.038	24.7	23.4
5	0.037	-30.8	0.039	30.8	-14.2
6	0.023	-34.3	0.016	34.3	62.6
7	0.025	-41.5	0.031	41.5	-79.6

Table A5.2 Harmonic Values of a Shallow Concave Trajectory

B A deeper arc or concave trajectory

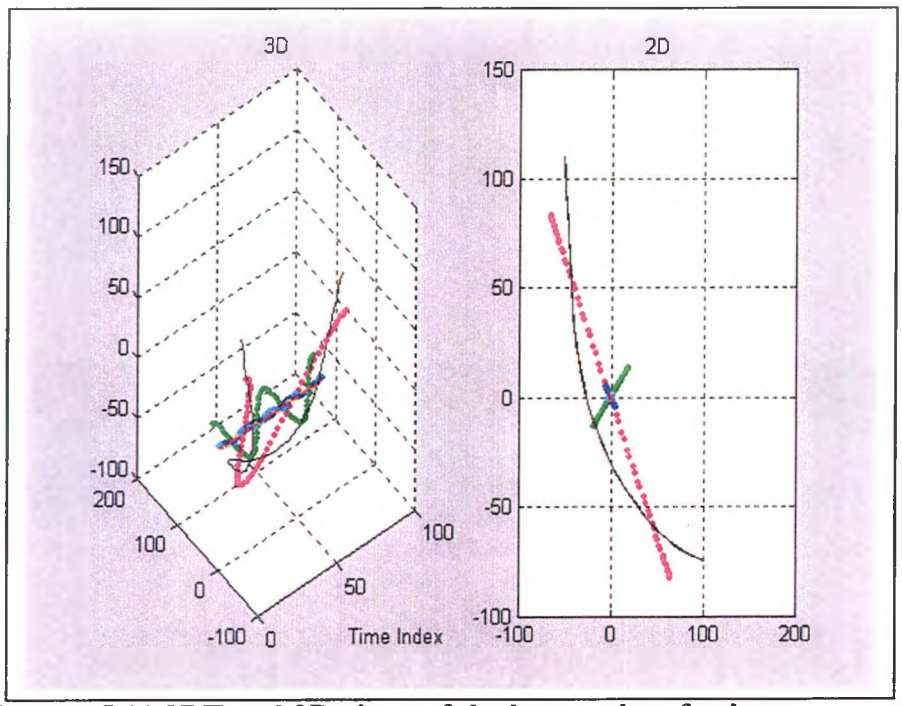


Figure A5.11 2DT and 2D views of the harmonics of a deep arc or concave trajectory.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-3.8	1.0	3.8	-128.3
2	0.219	-7.4	0.218	7.4	92.2
3	0.066	-11.0	0.067	11	-99.2
4	0.033	-14.2	0.031	14.2	117.8
5	0.041	-18.2	0.043	18.2	-128.4
6	0.021	-19.1	0.020	19.1	95.6
7	0.019	-23.6	0.022	23.6	-81.9

Table A5.3 Harmonic values of a deep arc or concave trajectory

C A Convex Trajectory

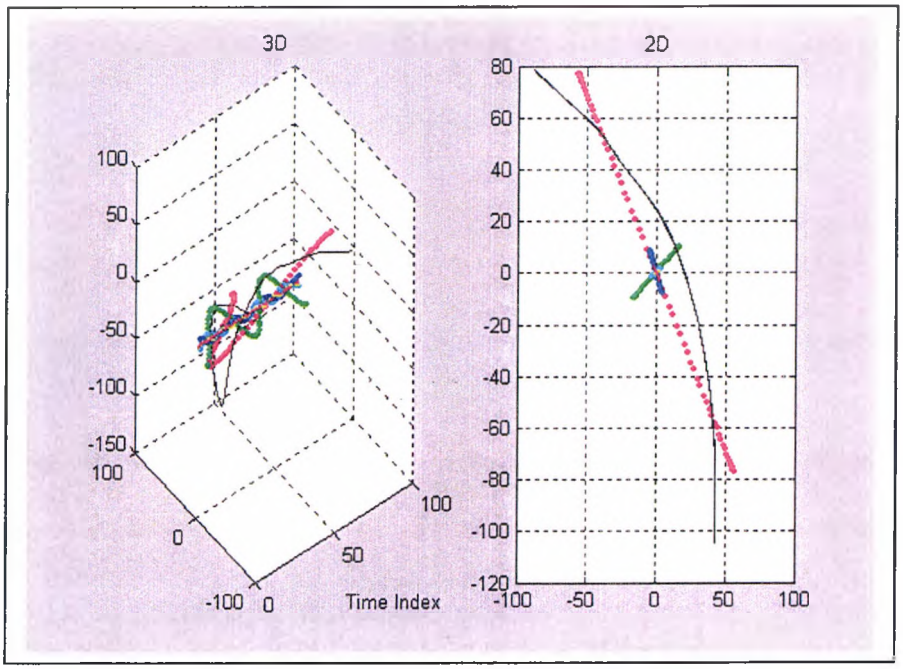


Figure A5.12 2DT and 2D views of the harmonics of a convex trajectory

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-4.5	1.00	4.5	-126.3
2	0.198	-9.0	0.198	-9.0	274.8
3	0.097	-13.7	0.097	13.7	-267.1
4	0.039	-18.4	0.038	18.4	273.4
5	0.043	-23.3	0.043	23.3	-280.9
6	0.029	151.9	0.028	-151.9	113.5
7	0.025	-33.7	0.026	33.7	-126.7

Table A5.4 Harmonic values of a deep convex trajectory

D An Elliptical Trajectory

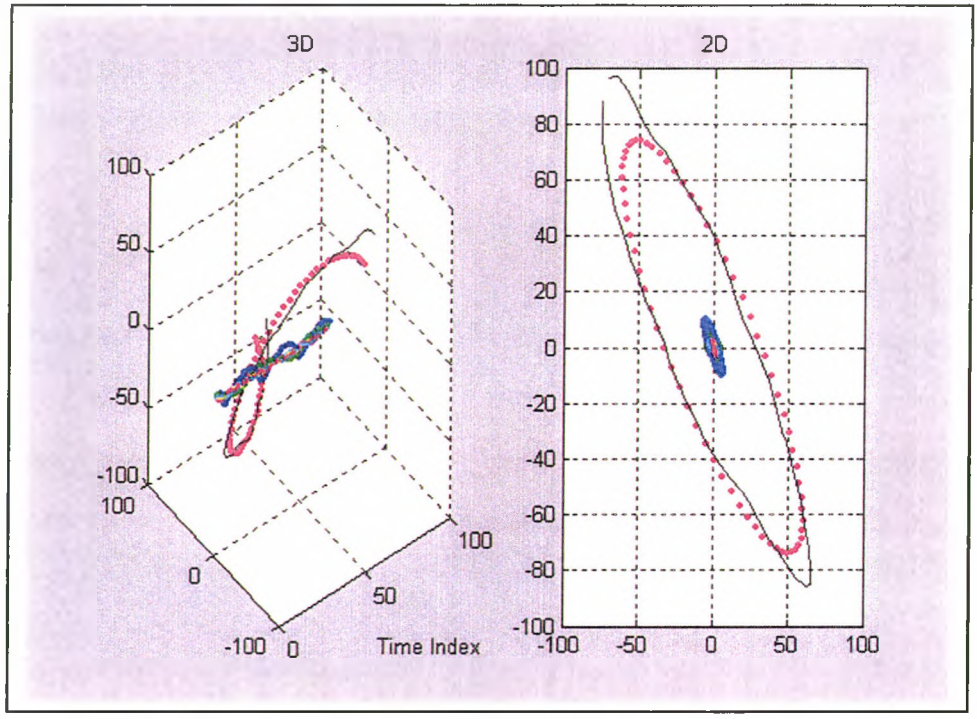


Figure A5.13 2DT and 2D views of an elliptical trajectory

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-11.1	0.555	11.1	-128.4
2	0.073	34.1	0.021	-34.1	-14.7
3	0.120	-41.3	0.074	41.3	22.9
4	0.029	14.3	0.018	-14.3	-14.9
5	0.046	-47.7	0.022	47.7	26.1
6	0.023	-51.7	0.001	51.7	56.8
7	0.032	110.0	0.020	47.7	26.1

Table A5.5 Harmonic values of an elliptic trajectory

E A 'Figure of Eight' Trajectory

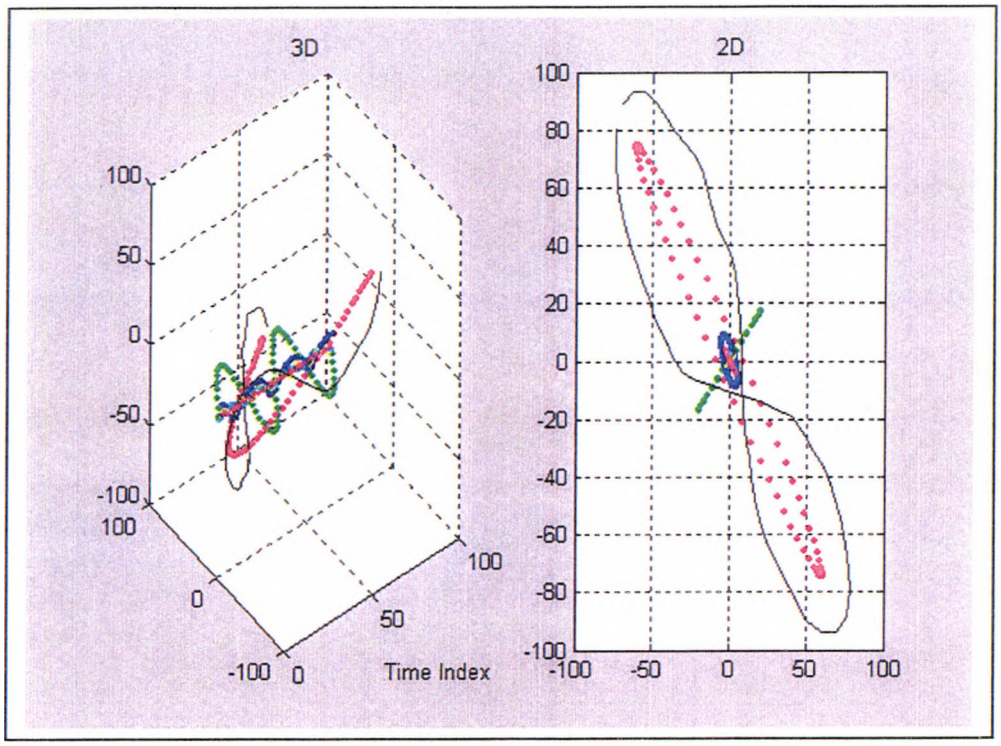


Figure A5.14 2DT and 2D views of a 'figure of eight' trajectory

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	1.98	1.156	-1.98	-129.3
2	0.299	90.0	0.291	-90	89.3
3	0.146	4.8	0.073	-4.8	-75.2
4	0.02	-26.1	0.02	26.1	221.0
5	0.057	1.0	0.0835	-1.0	-235.2
6	0.041	84.3	0.037	-84.3	97.3
7	0.035	-34	0.001	34	-1.8

Table A5.6 Harmonic values of 'figure of eight' trajectory

F An Oscillating Trajectory

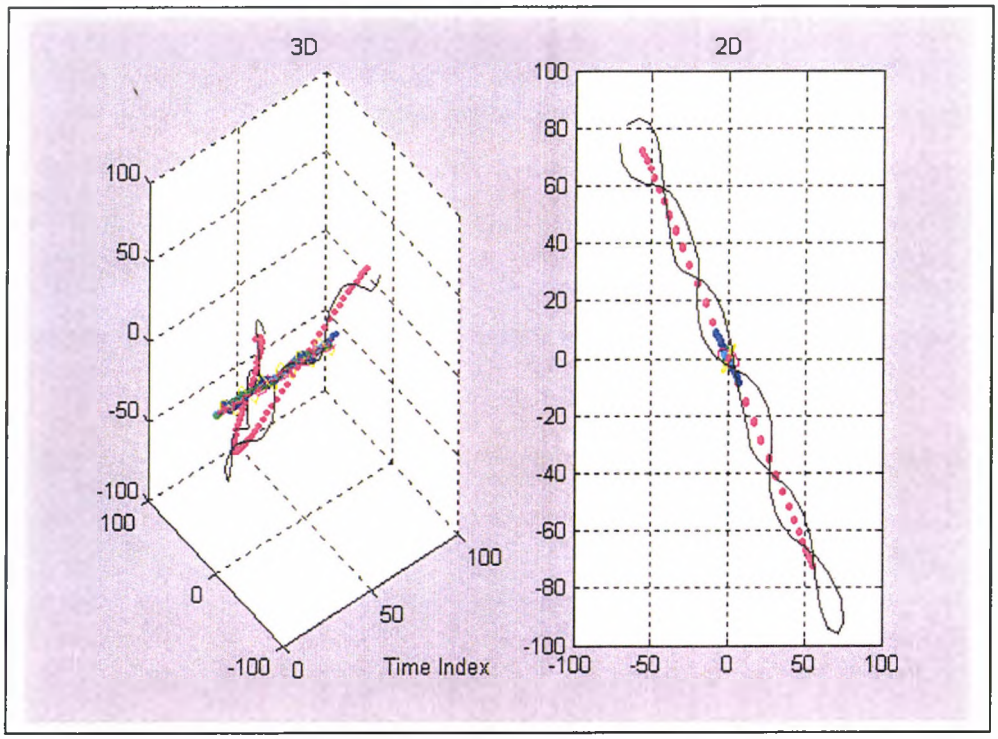


Figure A5.15 2DT and 2D views of the harmonics of an oscillatory trajectory

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	2.5	0.992	-2.5	-127.6
2	0.059	3.5	0.038	-3.5	182.4
3	0.129	-0.8	0.135	0.8	-185.3
4	0.023	-7.6	0.029	7.6	165.1
5	0.047	151.0	0.109	-151	-19.8
6	0.077	105.4	0.061	-105.4	-66.7
7	0.021	-45.8	0.020	45.8	-28.8

Table A5.7 Harmonic values of an oscillatory trajectory

A Real Gesture Trajectory

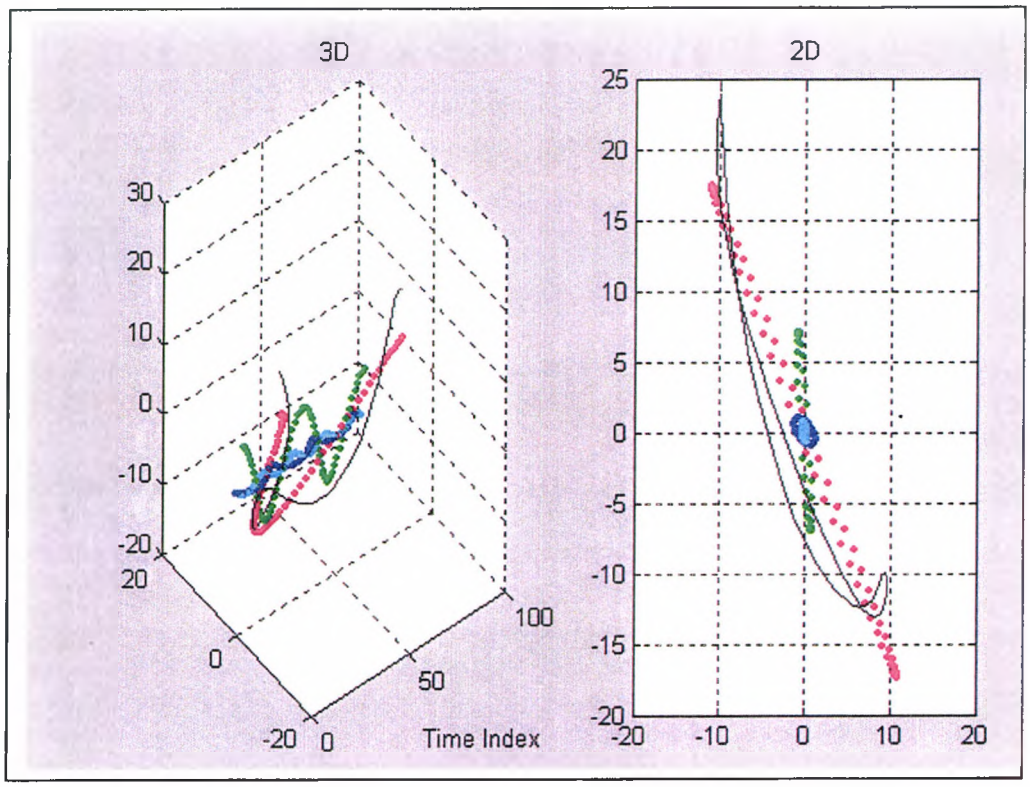


Figure A5.16 2DT and 2D Views of the harmonics of a real gesture trajectory

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	3	1.1	-3	-123
2	0.317	6	0.327	-6	26
3	0.042	27	0.103	-27	135
4	0.027	68	0.0513	-68	44

Table A5.8 Harmonic values of a real gesture trajectory

2 Examples of 'elliptic-corkscrews'

The third and fourth 'elliptic corkscrews' are shown in Figures A6.8 and Figures A6.9 respectively.

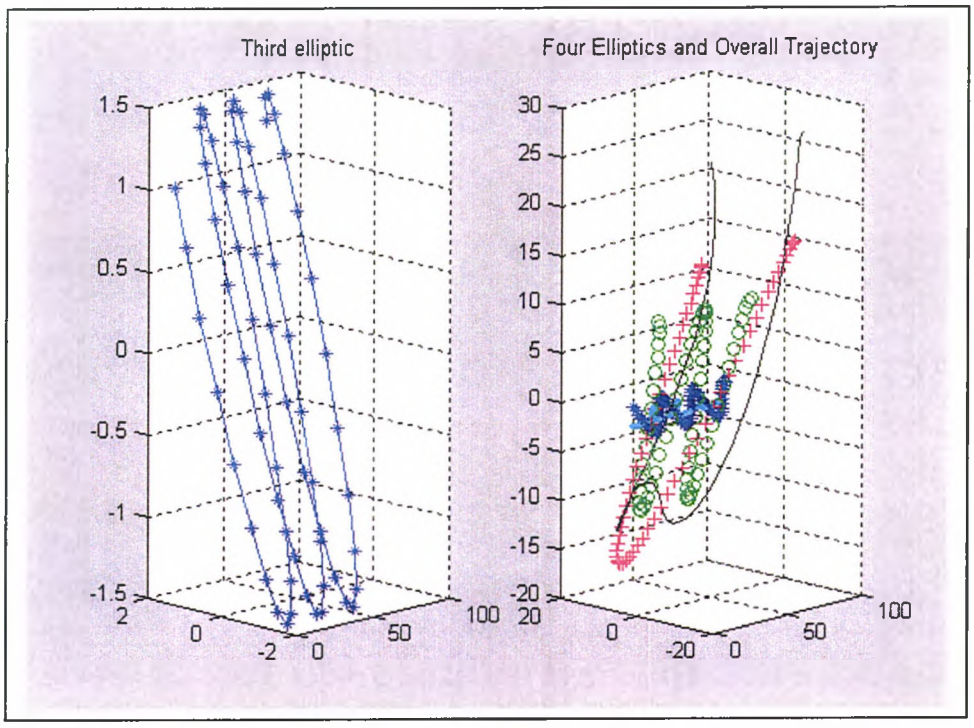


Figure A5.17 A third harmonic, single 'elliptic-corkscrew'

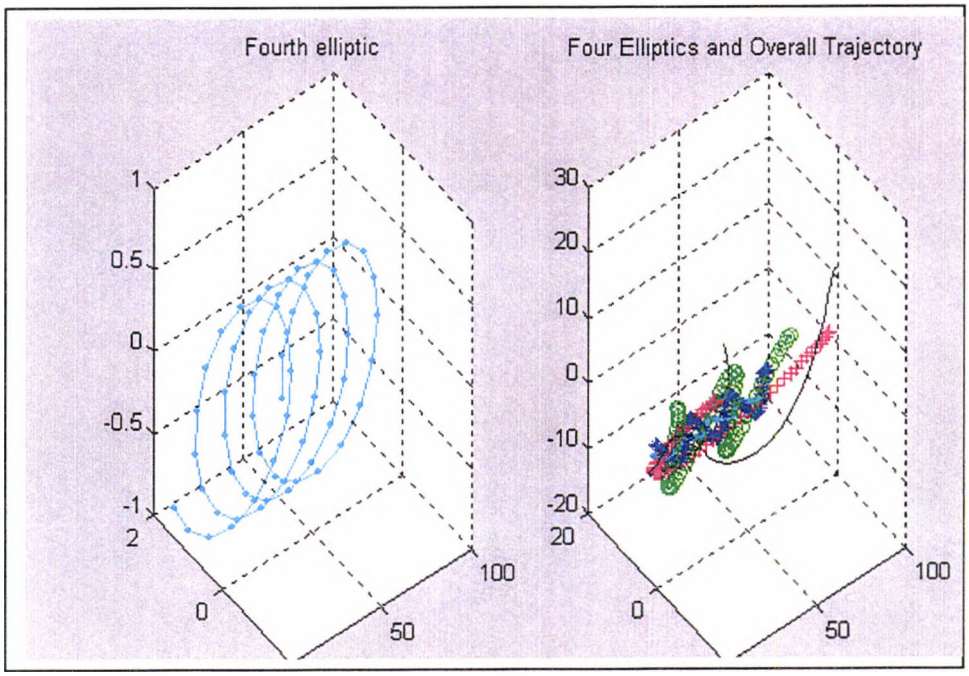


Figure A5.18 A fourth harmonic, single 'elliptic-corkscrew'

3 Truncation Performance

The stopping Tolerance set to 2, 5, 10 and 15, gave the following harmonic profiles respectively.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	16.7	1.229	-16.7	-16.9
2	0.456	2.5	0.361	-2.5	-27.9
3	0.321	-32.6	0.186	32.6	13.8
4	0.096	-17.8	0.039	17.8	-98.8
5	0.118	75.8	0.07	-75.8	132.4
6	0.117	-69.1	0.118	69.1	-31.5
7	0.042	-162.4	0.030	162.4	31.8

Table A5.9 Harmonics with stopping Tolerance 2; Gesture Length 37

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	22.7	1.23	-22.7	-15.4
2	0.450	14.3	0.374	-14.3	-28.1
3	0.348	-16.1	0.203	16.1	13.4
4	0.108	3.73	0.038	-3.73	-84.6
5	0.130	103.2	0.089	-103.2	111.7
6	0.107	-37.2	0.116	37.2	-27.8
7	0.061	-122.2	0.030	122.2	34.3

Table A5.10 Harmonics with stopping Tolerance 5; Gesture Length 36

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	35.0	1.211	-35.0	-12.5
2	0.435	39.7	0.383	-39.7	-27.9
3	0.391	18.6	0.223	-18.6	12.9
4	0.147	47.6	0.036	-47.6	-62.3
5	0.139	150.9	0.130	-150.9	75.0
6	0.106	40.9	0.112	-40.9	-15.9
7	0.097	-49.5	0.019	40.5	51.6

Table A5.11 Harmonics with stopping Tolerance 10; Gesture Length 34

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	49.8	1.331	-49.8	-12.6
2	0.444	63.5	0.505	-63.5	-25.0
3	0.467	58.0	0.401	-58.0	10.5
4	0.307	50.6	0.185	-50.6	-8.0
5	0.837	22.0	0.025	-22.0	-85.1
6	0.072	90.7	0.177	-90.7	88.2
7	0.155	68.3	0.170	-68.3	12.9

Table A5.12 Harmonics with stopping Tolerance 15; Gesture Length 32

4 OSA (Object Selection Algorithm) Performance

Data source -SCM objects

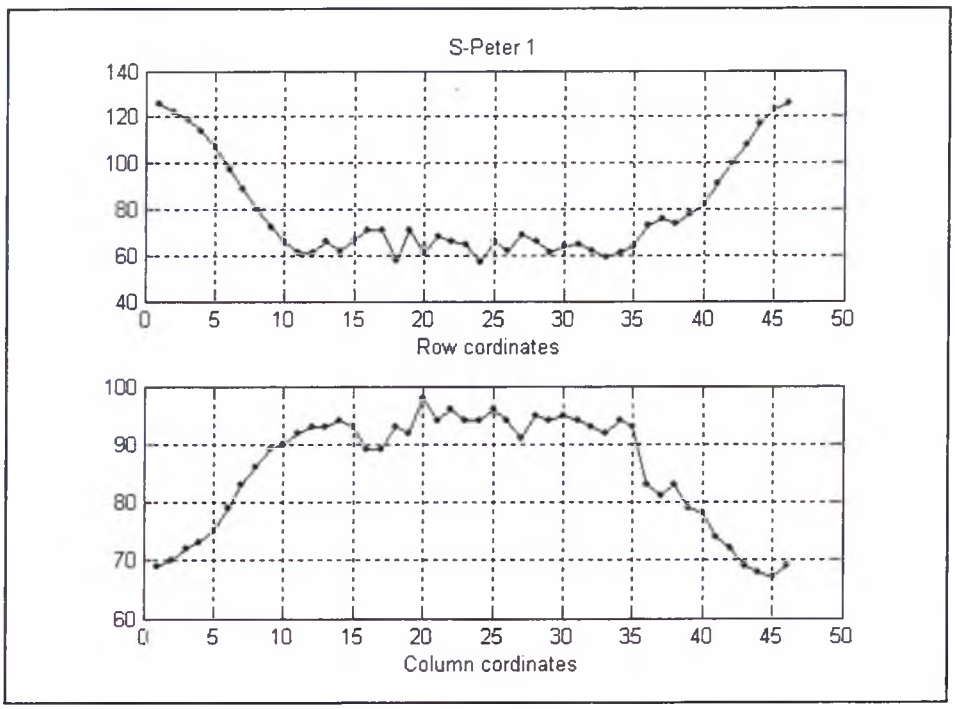


Figure A5.19 Row and column coordinates of a trajectory generated using SCM objects.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-2.5	0.888	2.5	26.6
2	0.618	-7.2	0.489	7.2	-8.5
3	0.245	7.7	0.209	-7.7	-7.0
4	0.050	77.6	0.096	-77.6	16.8
5	0.055	-68.3	0.040	68.3	4.8
6	0.058	-95.5	0.075	95.5	-24.7
7	0.020	4.7	0.040	-4.7	-108.1

Table A5.13 Harmonics generated using SCM objects.

Data source – SCMI objects

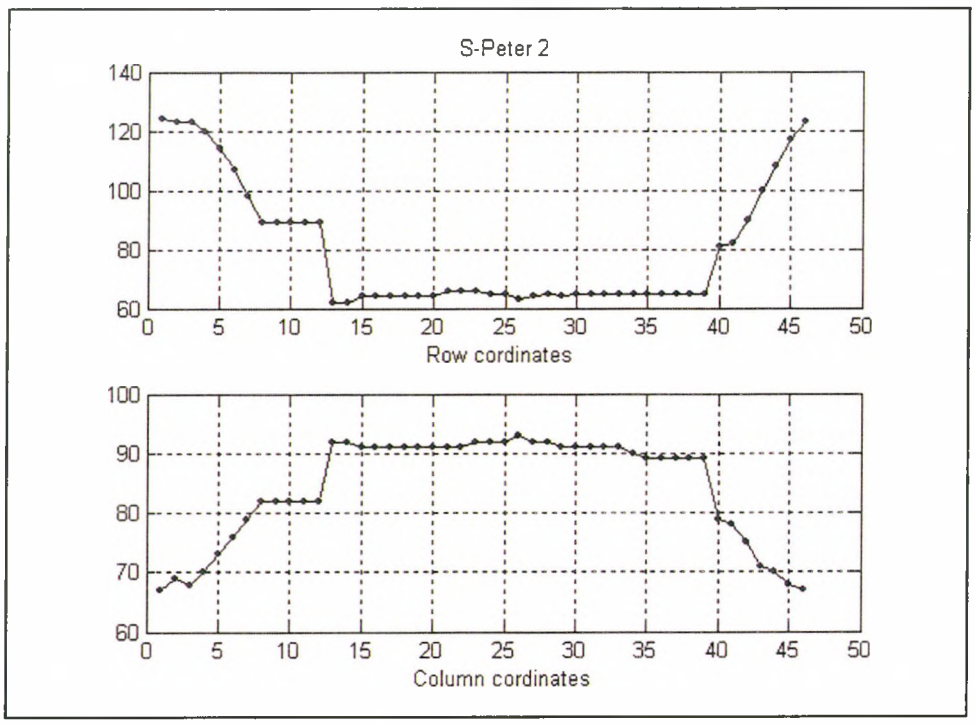


Figure A5.20 Row and column coordinates of a trajectory generated using SCMI objects.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	12.7	0.9	-12.7	23.3
2	0.549	23.0	0.451	-23.0	-5.0
3	0.175	8.2	0.125	-8.2	0.8
4	0.139	-41.3	0.095	41.3	0.4
5	0.056	26.1	0.058	-26.1	-7.7
6	0.057	115.3	0.073	-115.3	17.7
7	0.069	-2.4	0.045	2.4	-185.8

Table A5.14 Harmonics generated using SCMI objects.

Data source – Visually/Manually Recorded

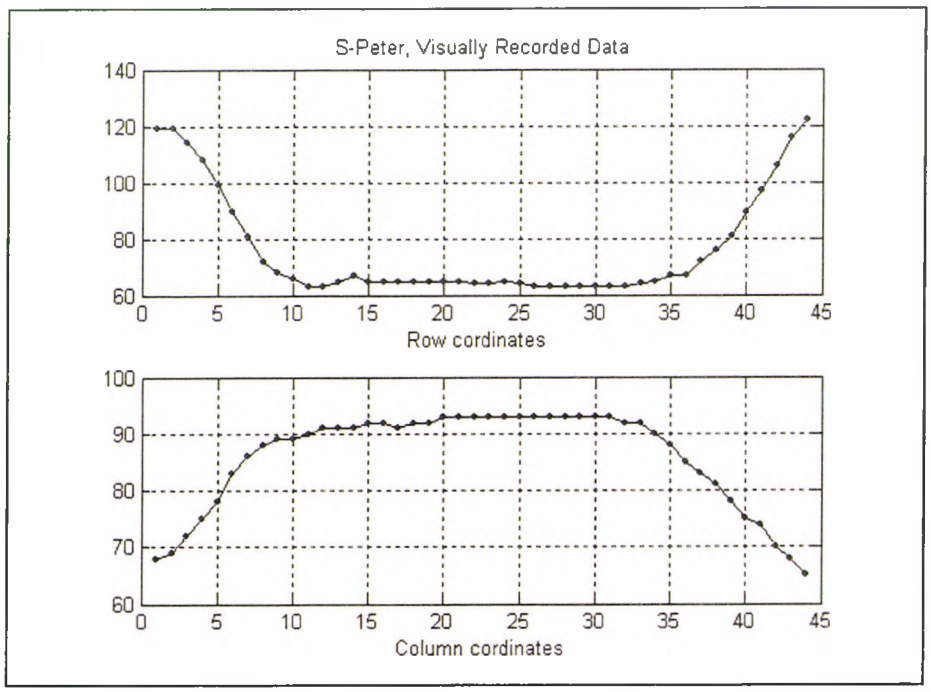


Figure A5.21 Row and column coordinates of a trajectory recorded visually.

Harmonic	Positive Magnitude, A_p	Positive phase ϕ	Negative Magnitude, A_n	Negative phase ϕ	Relative Orientation θ
1	1	-3.0	0.906	3.0	26.3
2	0.682	-9.4	0.546	9.4	-6.6
3	0.337	-3.2	0.263	3.2	-5.3
4	0.115	6	0.117	-6	0.4
5	0.025	-8.8	0.042	8.8	29.0
6	0.022	-67.2	0.036	67.2	-41.0
7	0.036	-64.2	0.019	64.2	4.2

Table A5.15 Harmonics generated from visually recorded coordinates

Appendix VI – Cluster Analysis - Matlab Toolbox (2006)

A Overview of K-Means Clustering

K-means clustering can best be described as a partitioning method. That is, the function *kmeans* partitions the observations in your data into K mutually exclusive clusters, and returns a vector of indices indicating to which of the k clusters it has assigned each observation. Unlike the hierarchical clustering methods used in linkage (see Hierarchical Clustering), *kmeans* does not create a tree structure to describe the groupings in your data, but rather creates a single level of clusters. Another difference is that K-means clustering uses the actual observations of objects or individuals in your data, and not just their proximities. These differences often mean that *kmeans* is more suitable for clustering large amounts of data.

kmeans treats each observation in your data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. You can choose from five different distance measures, depending on the kind of data you are clustering.

Each cluster in the partition is defined by its member objects and by its centroid, or centre. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized. *kmeans* computes cluster centroids differently for each distance measure, to minimize the sum with respect to the measure that you specify.

kmeans uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. You can control the details of the minimization using several optional input parameters to *kmeans*, including ones for the initial values of the cluster centroids, and for the maximum number of iterations.

B Hierarchical Clustering

Hierarchical clustering is a way to investigate grouping in your data, simultaneously over a variety of scales, by creating a cluster tree. The tree is not a single set of clusters, but rather a multilevel hierarchy, where clusters at one level are joined as clusters at the next higher level. This allows you to decide what level or scale of clustering is most appropriate in your application.

To perform hierarchical cluster analysis on a data set using the Statistics Toolbox functions, follow this procedure:

1 **Find the similarity or dissimilarity between every pair of objects in the data set.** In this step, you calculate the distance between objects using the *pdist* function. The *pdist* function supports many different ways to compute this measurement. See Finding the Similarities Between Objects for more information.

2 **Group the objects into a binary, hierarchical cluster tree**. In this step, you link pairs of objects that are in close proximity using the linkage function. The linkage function uses the distance information generated in step 1 to determine the proximity of objects to each other. As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed. See Defining the Links Between Objects for more information.

3 **Determine where to cut the hierarchical tree into clusters**. In this step, you use the cluster function to prune branches off the bottom of the hierarchical tree, and assign all the objects below each cut to a single cluster. This creates a partition of the data. The cluster function can create these clusters by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point

4 Finding the Similarities Between Objects

You use the *pdist* function to calculate the distance between every pair of objects in a data set. For a data set made up of m objects, there are pairs in the data set. The result of this computation is commonly known as a distance or dissimilarity matrix.

There are many ways to calculate this distance information. By default, the *pdist* function calculates the Euclidean distance between objects; however, you can specify one of several other options. See *pdist* for more information.

For example, consider a data set, X, made up of five objects where each object is a set of x, y coordinates.

Object 1: 1, 2

Object 2: 2.5, 4.5

Object 3: 2, 2

Object 4: 4, 1.5

Object 5: 4, 2.5

You can define this data set as a matrix

X = [1 2; 2.5 4.5; 2 2;4 1.5; 4 2.5]

and pass it to *pdist*. The pdist function calculates the distance between object 1 and object 2, object 1 and object 3, and so on until the distances between all the pairs have been calculated. The following figure plots these objects in a graph. The Euclidean distance between object 2 and object 3 is shown to illustrate one interpretation of distance.

Returning Distance Information. The *pdist* function returns this distance information in a vector, Y, where each element contains the distance between a pair of objects.

Y = *pdist*(X)

Y =

Columns 1 through 7

2.9155 1.0000 3.0414 3.0414 2.5495 3.3541 2.5000

Columns 8 through 10

2.0616 2.0616 1.0000

To make it easier to see the relationship between the distance information generated by *pdist* and the objects in the original data set, you can reformat the distance vector into a matrix using the *squareform* function. In this matrix, element i,j corresponds to the distance between object i and object j in the original data set. In the following example, element 1,1 represents the distance between object 1 and itself (which is zero). Element 1,2 represents the distance between object 1 and object 2, and so on.

Squareform (Y)

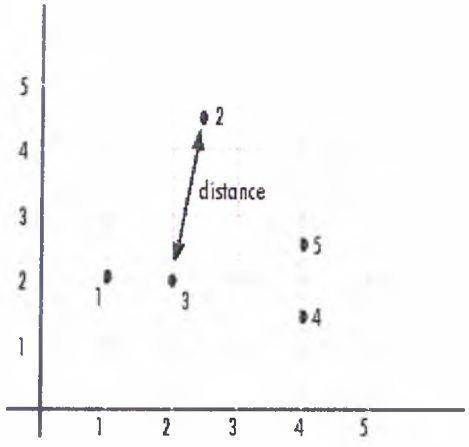
ans =

0	2.9155	1.0000	3.0414	3.0414
2.9155	0	2.5495	3.3541	2.5000
1.0000	2.5495	0	2.0616	2.0616
3.0414	3.3541	2.0616	0	1.0000
3.0414	2.5000	2.0616	1.0000	0

Defining the Links Between Objects

Once the proximity between objects in the data set has been computed, you can determine how objects in the data set should be grouped into clusters, using the linkage function. The linkage function takes the distance information generated by *pdist* and links pairs of objects that are close together into binary clusters (clusters made up of two objects). The linkage function then links these newly formed clusters to each other and to other objects to create bigger clusters until all the objects in the original data set are linked together in a hierarchical tree.

For example, given the distance vector Y generated by *pdist* from the sample data set of x- and y-coordinates, the linkage function generates a hierarchical cluster tree, returning the linkage information in a matrix, Z.



Z = *linkage*(Y)

Z =

4.0000	5.0000	1.0000
1.0000	3.0000	1.0000
6.0000	7.0000	2.0616
2.0000	8.0000	2.5000

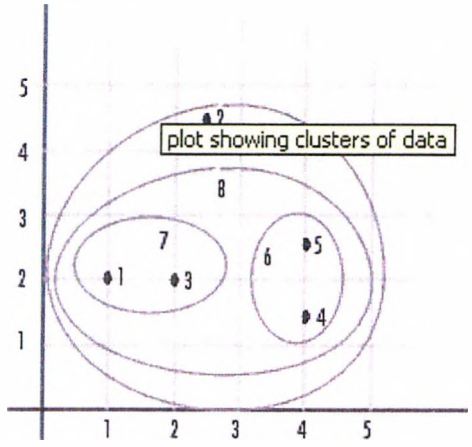
In this output, each row identifies a link between objects or clusters. The first two columns identify the objects that have been linked, that is, object 1, object 2, and so on. The third column contains the distance between these objects. For the sample data set of x- and y-coordinates, the linkage function begins by grouping objects 1 and 3, which have the closest proximity (distance value = 1.0000). The linkage function continues by grouping objects 4 and 5, which also have a distance value of 1.0000.

The third row indicates that the linkage function grouped objects 6 and 7. If the original sample data set contained only five objects, what are objects 6 and 7? Object 6 is the newly formed binary cluster created by the grouping of objects 1 and 3. When the linkage function groups two objects into a new cluster, it must assign the cluster a unique index value, starting with the value m+1, where m is the number of objects in the original data set. (Values 1 through m are already used by the original data set.) Similarly, object 7 is the cluster formed by grouping objects 4 and 5.

linkage uses distances to determine the order in which it clusters objects. The distance vector Y contains the distances between the original objects 1 through 5.

But linkage must also be able to determine distances involving clusters that it creates, such as objects 6 and 7. By default, linkage uses a method known as single linkage. However, there are a number of different methods available. See the linkage reference page for more information.

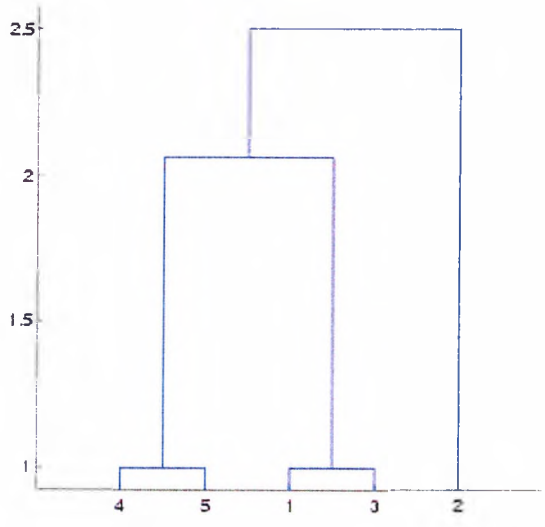
As the final cluster, the linkage function grouped object 8, the newly formed cluster made up of objects 6 and 7, with object 2 from the original data set. The following figure graphically illustrates the way linkage groups the objects into a hierarchy of clusters.



Plotting the Cluster Tree

The hierarchical, binary cluster tree created by the linkage function is most easily understood when viewed graphically. The Statistics Toolbox includes the dendrogram function that plots this hierarchical tree information as a graph, as in the following example.

dendrogram (Z)



In the figure, the numbers along the horizontal axis represent the indices of the objects in the original data set. The links between objects are represented as upside-down U-shaped lines. The height of the U indicates the distance between the objects. For example, the link representing the cluster containing objects 1 and 3 has a height of 1. The link representing the cluster that groups object 2 together with objects 1, 3, 4, and 5, (which are already clustered as object 8) has a height of 2.5. The height represents the distance linkage computes between objects 2 and 8. For more information about creating a dendrogram diagram, see the dendrogram reference page.

Evaluating Cluster Formation

After linking the objects in a data set into a hierarchical cluster tree, you might want to verify that the distances (that is, heights) in the tree reflect the original distances accurately. In addition, you might want to investigate natural divisions that exist among links between objects. The Statistics Toolbox provides functions to perform both these tasks, as described in the following sections:

Verifying the Cluster Tree

Getting More Information About Cluster Links

Verifying the Cluster Tree. In a hierarchical cluster tree, any two objects in the original data set are eventually linked together at some level. The height of the link represents the distance between the two clusters that contain those two objects. This height is known as the *cophenetic* distance between the two objects. One way to measure how well the cluster tree generated by the linkage function reflects your data is to compare the *cophenetic* distances with the original distance data generated by the *pdist* function. If the clustering is valid, the linking of objects in the cluster tree should have a strong correlation with the distances between objects in the distance vector. The *cophenet* function compares these two sets of values and computes their correlation, returning a value called the *cophenetic* correlation coefficient. The closer the value of the *cophenetic* correlation coefficient is to 1, the more accurately the clustering solution reflects your data.

You can use the *cophenetic* correlation coefficient to compare the results of clustering the same data set using different distance calculation methods or clustering algorithms. For example, you can use the *cophenet* function to evaluate the clusters created for the sample data set

c = *cophenet*(Z, Y)

c =

0.8615

where Z is the matrix output by the linkage function and Y is the distance vector output by the *pdist* function.

Execute *pdist* again on the same data set, this time specifying the city block metric. After running the linkage function on this new pdist output using the average linkage method, call cophenet to evaluate the clustering solution.

```
Y = pdist(X,'cityblock');

Z = linkage(Y,'average');

c = cophenet(Z, Y)

c =

   0.9044
```

The cophenetic correlation coefficient shows that using a different distance and linkage method creates a tree that represents the original distances slightly better.

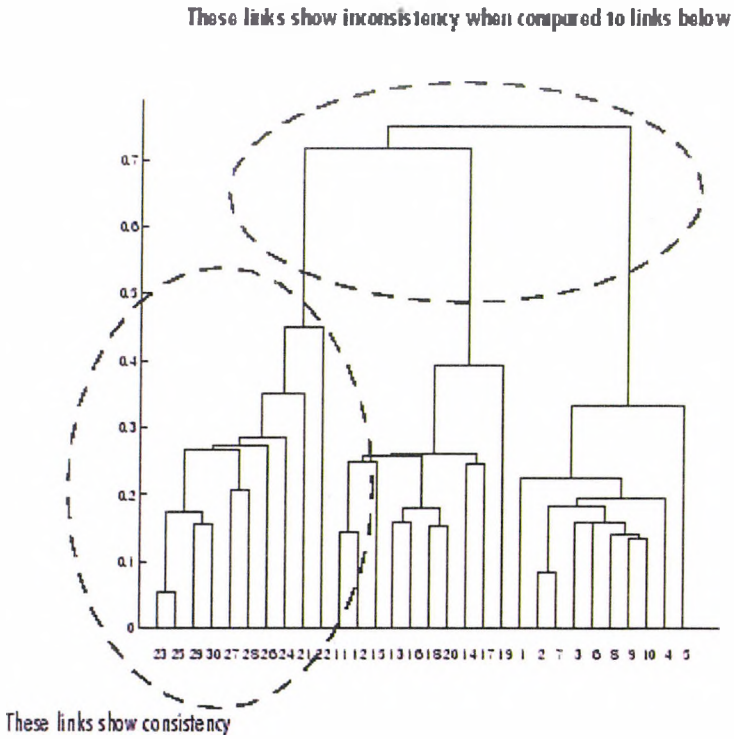
Getting More Information About Cluster Links. One way to determine the natural cluster divisions in a data set is to compare the height of each link in a cluster tree with the heights of neighbouring links below it in the tree.

A link that is approximately the same height as the links below it indicates that there are no distinct divisions between the objects joined at this level of the hierarchy. These links are said to exhibit a high level of consistency, because the distance between the objects being joined is approximately the same as the distances between the objects they contain.

On the other hand, a link whose height differs noticeably from the height of the links below it indicates that the objects joined at this level in the cluster tree are much farther apart from each other than their components were when they were joined. This link is said to be inconsistent with the links below it.

In cluster analysis, inconsistent links can indicate the border of a natural division in a data set. The cluster function uses a quantitative measure of inconsistency to determine where to partition your data set into clusters.

The following dendrogram, created using a data set of random numbers, illustrates inconsistent links. Note how the objects in the dendrogram fall into three groups that are connected by links at a much higher level in the tree. These links are inconsistent when compared with the links below them in the hierarchy.



The relative consistency of each link in a hierarchical cluster tree can be quantified and expressed as the inconsistency coefficient. This value compares the height of a link in a cluster hierarchy with the average height of links below it. Links that join distinct clusters have a low inconsistency coefficient; links that join indistinct clusters have a high inconsistency coefficient.

To generate a listing of the inconsistency coefficient for each link in the cluster tree, use the inconsistent function. By default, the inconsistent function compares each link in the cluster hierarchy with adjacent links that are less than two levels below it in the cluster hierarchy. This is called the depth of the comparison. You can also specify other depths. The objects at the bottom of the cluster tree, called leaf nodes, that have no further objects below them, have an inconsistency coefficient of zero. Clusters that join two leaves also have a zero inconsistency coefficient.

For example, you can use the inconsistent function to calculate the inconsistency values for the links created by the linkage function in Defining the Links Between Objects.

I = inconsistent(Z)

I =

1.0000	0	1.0000	0
1.0000	0	1.0000	0
1.3539	0.6129	3.0000	1.1547
2.2808	0.3100	2.0000	0.7071

The inconsistent function returns data about the links in an (m-1)-by-4 matrix, whose columns are described in the following table.

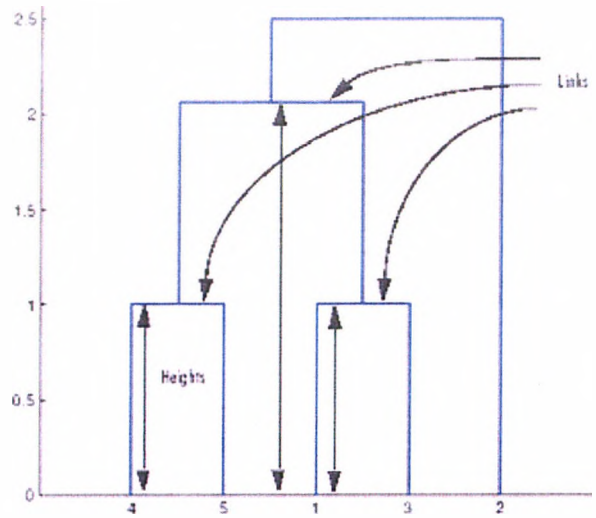
Column	Description
1	Mean of the heights of all the links included in the calculation
2	Standard deviation of all the links included in the calculation
3	Number of links included in the calculation
4	Inconsistency coefficient

In the sample output, the first row represents the link between objects 4 and 5. This cluster is assigned the index 6 by the linkage function. Because both 4 and 5 are leaf nodes, the inconsistency coefficient for the cluster is zero. The second row represents the link between objects 1 and 3, both of which are also leaf nodes. This cluster is assigned the index 7 by the linkage function.

The third row evaluates the link that connects these two clusters, objects 6 and 7. (This new cluster is assigned index 8 in the linkage output). Column 3 indicates that three links are considered in the calculation: the link itself and the two links directly below it in the hierarchy. Column 1 represents the mean of the heights of these links. The inconsistent function uses the height information output by the linkage function to calculate the mean. Column 2 represents the standard deviation between the links. The last column contains the inconsistency value for these links, 1.1547. It is the difference between the current link height and the mean, normalized by the standard deviation:

```
>> (2.0616 - 1.3539) / .6129

ans =

   1.1547
```

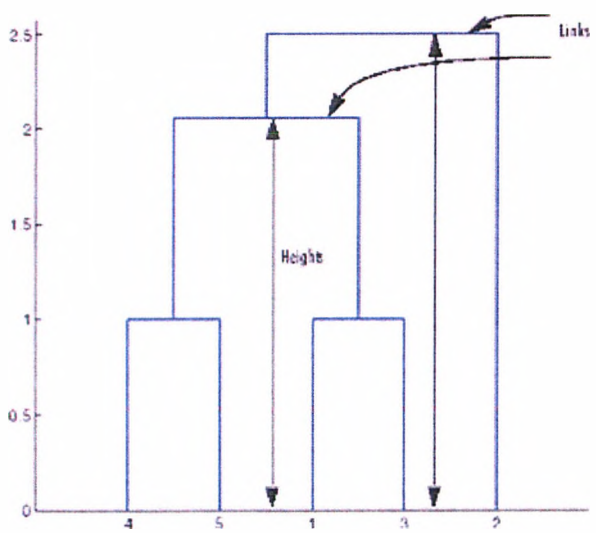
The following figure illustrates the links and heights included in this calculation.



Note In the preceding figure, the lower limit on the y-axis is set to 0 to show the heights of the links. To set the lower limit to 0, select Axes Properties from the Edit menu, click the Y Axis tab, and enter 0 in the field immediately to the right of Y Limits.

Row 4 in the output matrix describes the link between object 8 and object 2. Column 3 indicates that two links are included in this calculation: the link itself and the link directly below it in the hierarchy. The inconsistency coefficient for this link is 0.7071.

The following figure illustrates the links and heights included in this calculation.



Creating Clusters

After you create the hierarchical tree of binary clusters, you can prune the tree to partition your data into clusters using the cluster function. The cluster function lets you create clusters in two ways, as discussed in the following sections:

Finding Natural Divisions in Data

Specifying Arbitrary Clusters

Finding Natural Divisions in Data. The hierarchical cluster tree may naturally divide the data into distinct, well-separated clusters. This can be particularly evident in a dendrogram diagram created from data where groups of objects are densely packed in certain areas and not in others. The inconsistency coefficient of the links in the cluster tree can identify these divisions where the similarities between objects change abruptly. (See Evaluating Cluster Formation for more information about the inconsistency coefficient.) You can use this value to determine where the cluster function creates cluster boundaries.

For example, if you use the cluster function to group the sample data set into clusters, specifying an inconsistency coefficient threshold of 1.2 as the value of the cutoff argument, the cluster function groups all the objects in the sample data set into one cluster. In this case, none of the links in the cluster hierarchy had an inconsistency coefficient greater than 1.2.

```
T = cluster(Z,'cutoff',1.2)

T =

     1
     1
     1
     1
     1
```

The cluster function outputs a vector, T, that is the same size as the original data set. Each element in this vector contains the number of the cluster into which the corresponding object from the original data set was placed.

If you lower the inconsistency coefficient threshold to 0.8, the cluster function divides the sample data set into three separate clusters.

```
T = cluster(Z,'cutoff', 0.8)

T =

     1
     3
     1
     2
     2
```

This output indicates that objects 1 and 3 were placed in cluster 1, objects 4 and 5 were placed in cluster 2, and object 2 was placed in cluster 3.

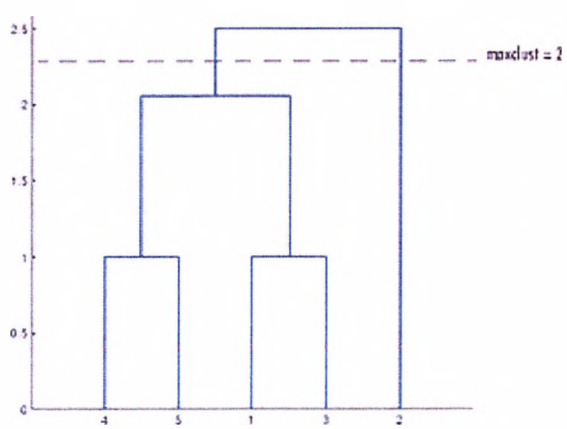
When clusters are formed in this way, the cutoff value is applied to the inconsistency coefficient. These clusters may, but do not necessarily, correspond to a horizontal slice across the dendrogram at a certain height. If you want clusters corresponding to a horizontal slice of the dendrogram, you can either use the criterion option to specify that the cutoff should be based on distance rather than inconsistency, or you can specify the number of clusters directly as described in the following section.

Specifying Arbitrary Clusters. Instead of letting the cluster function create clusters determined by the natural divisions in the data set, you can specify the number of clusters you want created.

For example, you can specify that you want the cluster function to partition the sample data set into two clusters. In this case, the cluster function creates one cluster containing objects 1, 3, 4, and 5 and another cluster containing object 2.

```
T = cluster (Z,'maxclust', 2)

T =

     2
     1
     2
     2
     2
```

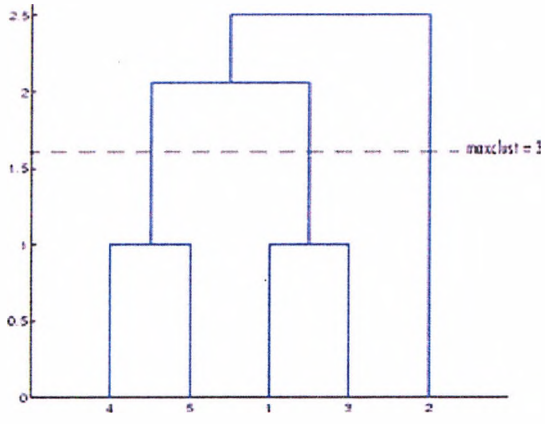
To help you visualize how the cluster function determines these clusters, the following figure shows the dendrogram of the hierarchical cluster tree. The horizontal dashed line intersects two lines of the dendrogram, corresponding to setting 'maxclust' to 2. These two lines partition the objects into two clusters: the objects below the left-hand line, namely 1, 3, 4, and 5, belong to one cluster, while the object below the right-hand line, namely 2, belongs to the other cluster.



On the other hand, if you set '*maxclust*' to 3, the cluster function groups objects 4 and 5 in one cluster, objects 1 and 3 in a second cluster, and object 2 in a third cluster. The following command illustrates this.

```
T = cluster(Z,'maxclust',3)

T =

   1

   3

   1

   2

   2
```

This time, the cluster function cuts off the hierarchy at a lower point, corresponding to the horizontal line that intersects three lines of the dendrogram in the following figure.



3 Distance Metrics

Y = pdist(X,distance) computes the distance between objects in the data matrix, X, using the method specified by distance, where distance can be any of the following character strings that identify ways to compute the distance.

'euclidean' Euclidean distance (default)

'seuclidean' Standardized Euclidean distance. Each coordinate in the sum of squares is inverse weighted by the sample variance of that coordinate.

'mahalanobis' Mahalanobis distance

'cityblock' City Block metric

'minkowski' Minkowski metric

'cosine' One minus the cosine of the included angle between points (treated as vectors)

'correlation' One minus the sample correlation between points (treated as sequences of values).

'spearman' One minus the sample Spearman's rank correlation between observations, treated as sequences of values

'hamming' Hamming distance, the percentage of coordinates that differ

'jaccard' One minus the Jaccard coefficient, the percentage of nonzero coordinates that differ

'chebychev' Chebychev distance (maximum coordinate difference)

Mathematical Definitions of Methods. Given an *m*-by-*n* data matrix X, which is treated as *m (1-by-n)* row vectors x_1, x_2,..., x_m, the various distances between the vector x_r and x_s are defined as follows:

- Euclidean distance:

$$d_{rs}^2 = (x_r - x_s)(x_r - x_s)'$$

- Standardized Euclidean distance:

$$d_{rs}^2 = (x_r - x_s)D^{-1}(x_r - x_s)'$$

where D is the diagonal matrix with diagonal elements given by v_j^2, which denotes the variance of the variable X_j over the m objects.

- Mahalanobis distance:

$$d_{rs}^2 = (x_r - x_s)'V^{-1}(x_r - x_s)$$

where V is the sample covariance matrix.

- City Block metric:

$$d_{rs} = \sum_{j=1}^{n} |x_{rj} - x_{sj}|$$

- Minkowski metric:

$$d_{rs} = \left\{ \sum_{j=1}^{n} |x_{rj} - x_{sj}|^p \right\}^{1/p}$$

Notice that when $p = 1$, it is the City Block case, and when $p = 2$, it is the Euclidean case.

Note: not all definitions are given here.

4 Linkage Methods

Z = linkage(Y) creates a hierarchical cluster tree, using the Single Linkage algorithm. The input matrix, Y, is the distance vector output by the pdist function, a vector of length $(m-1) \cdot m/2$ by 1, where m is the number of objects in the original dataset.

Z = linkage(Y,'*method*') computes a hierarchical cluster tree using the algorithm specified by '*method*'. method can be any of the following character strings that identify ways to create the cluster hierarchy.

String	Meaning
'single'	Shortest distance (default)
'complete'	Largest distance
'average'	Average distance
'centroid'	Centroid distance
'ward'	Incremental sum of squares

Note not all linkage methods are given here.

The output, Z, is an m-1 by 3 matrix containing cluster tree information. The leaf nodes in the cluster hierarchy are the objects in the original dataset, numbered from 1 to m. They are the singleton clusters from which all higher clusters are built. Each newly formed cluster, corresponding to row i in Z, is assigned the index $m+i$, where m is the total number of initial leaves.

Columns 1 and 2, Z(i,1:2), contain the indices of the objects that were linked in pairs to form a new cluster. This new cluster is assigned the index value $m+i$. There are $m-1$ higher clusters that correspond to the interior nodes of the hierarchical cluster tree.

Column 3, Z(i,3), contains the corresponding linkage distances between the objects paired in the clusters at each row i.

For example, consider a case with 30 initial nodes. If the tenth cluster formed by the linkage function combines object 5 and object 7 and their distance is 1.5, then row 10 of Z will contain the values (5,7,1.5). This newly formed cluster will have the index 10+30=40. If cluster 40 shows up in a later row, that means this newly formed cluster is being combined again into some bigger cluster.

Mathematical Definitions. The '*method*' argument is a character string that specifies the algorithm used to generate the hierarchical cluster tree information. These linkage algorithms are based on various measurements of proximity between two groups of objects. If n_r is the number of objects in cluster r and n_s is the number of objects in cluster s, and x_{ri} is the ith object in cluster r, the definitions of these various measurements are as follows:

- *Single linkage*, also called *nearest neighbor*, uses the smallest distance between objects in the two groups.

$$d(r,s) = min(dist(x_{ri},x_{sj})), i \in (i,\ldots,n_r), j \in (1,\ldots,n_s)$$

- *Complete linkage*, also called *furthest neighbor*, uses the largest distance between objects in the two groups.

$$d(r,s) = max(dist(x_{ri},x_{sj})), i \in (1,\ldots,n_r), j \in (1,\ldots,n_s)$$

- *Average linkage* uses the average distance between all pairs of objects in cluster r and cluster s.

$$d(r,s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} dist(x_{ri},x_{sj})$$

- *Centroid linkage* uses the distance between the centroids of the two groups

$$d(r,s) = d(\bar{x}_r, \bar{x}_s)$$

where:

$$\bar{x}_r = \frac{1}{n_r} \sum_{i=1}^{n_r} x_{ri}$$

and $\bar{x}_s$ is defined similarly.

- *Ward linkage* uses the incremental sum of squares; that is, the increase in the total within-group sum of squares as a result of joining groups r and s. It is given by

$$d(r,s) = n_r n_s d_{rs}^2 / (n_r + n_s)$$

where d_{rs}^2 is the distance between cluster r and cluster s defined in the Centroid linkage. The within-group sum of squares of a cluster is defined as the sum of the squares of the distance between all objects in the cluster and the centroid of the clus

5 Linkage Examples

Reference: Matlab Statistics toolbox

Linking and distance metrics

A data set X is made up of five objects where each object is a set of x, y coordinates:

Object No	'x' coordinate	'y' coordinate
1	1	2
2	1.5	4.5
3	2	2
4	4	1.5
5	4	2.5

Table A6.1 Object number and coordinate value

The function 'pdist' returns distance (Euclid) information as a vector or matrix, Y.

0	2.9155	1	3.01414	3.01414
2.9155	0	2.5495	3.3541	2.5
1	2.5495	0	2.0616	2.0616
3.01414	3.3541	2.0616	0	1
3.01414	2.5	2.0616	1	0

Table A6.2 Euclidean distance matrix derived from the five objects

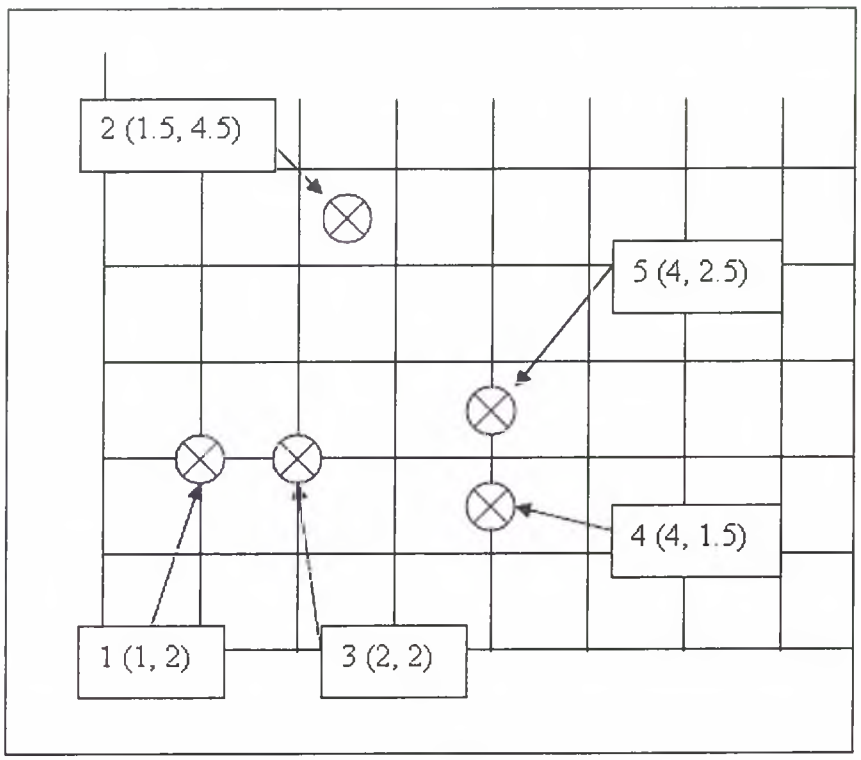


Figure A6.1 Diagram showing object number and coordinates for the five objects used in the clustering examples to show the differences between distance metrics and linkage methods.

Euclidean distance metric

'Single' linking

The linkage function takes the distance information and links together pairs of objects that are close to each other into binary clusters. The 'linkage' cluster then links together these newly formed clusters to each other to form bigger clusters.

Object No.	Object No	Distance apart
4	5	1
1	3	1
6	7	2.0616
2	8	2.5

Table A6.3 Distances of objects using 'single' linking and Euclidean distance metric

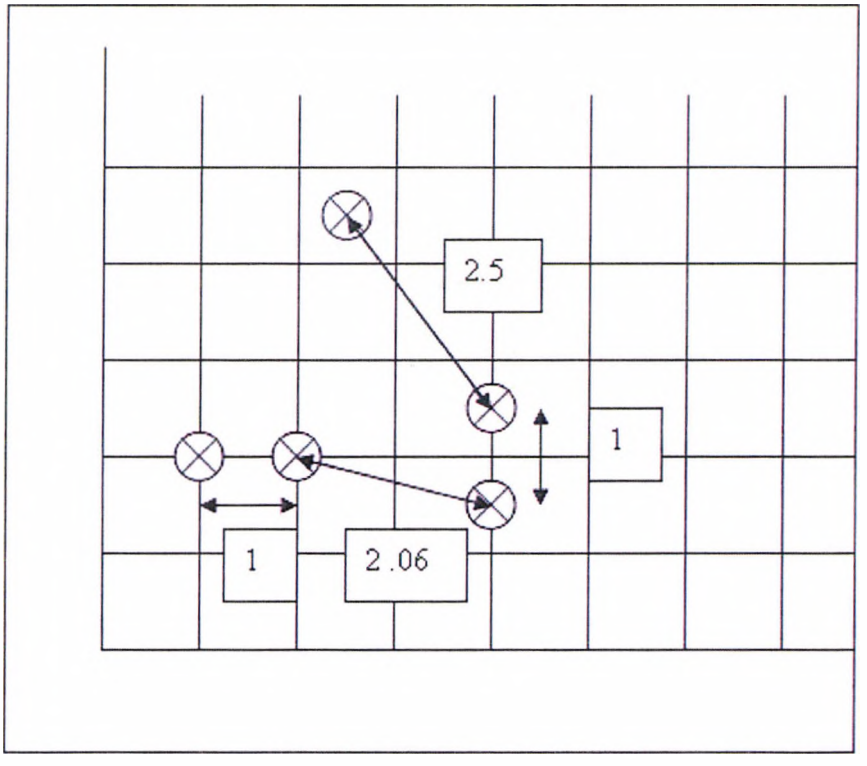


Figure A6.2 Distances between objects for 'single' linking and Euclidean distance metric

Euclidean distance measure and 'single' or 'nearest neighbour' linkage

The above diagram shows the distance generated in the linkage matrix. This gives rise to the dendrogram as shown below: -

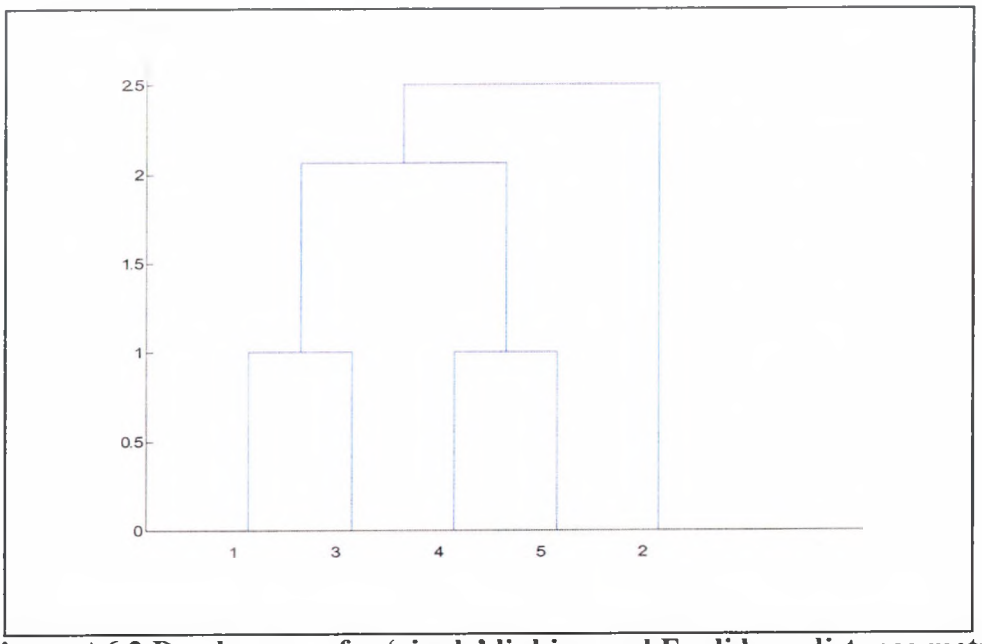


Figure A6.3 Dendrogram for 'single' linking and Euclidean distance metric

Complete linking

The 'complete or furthest distance linking gives: -

Object No.	Object No	Distance apart
1	3	1
4	5	1
6	2	2.9155
8	7	3.3541

Table A6.4 Distances of objects using 'complete' linking and Euclidean distance metric

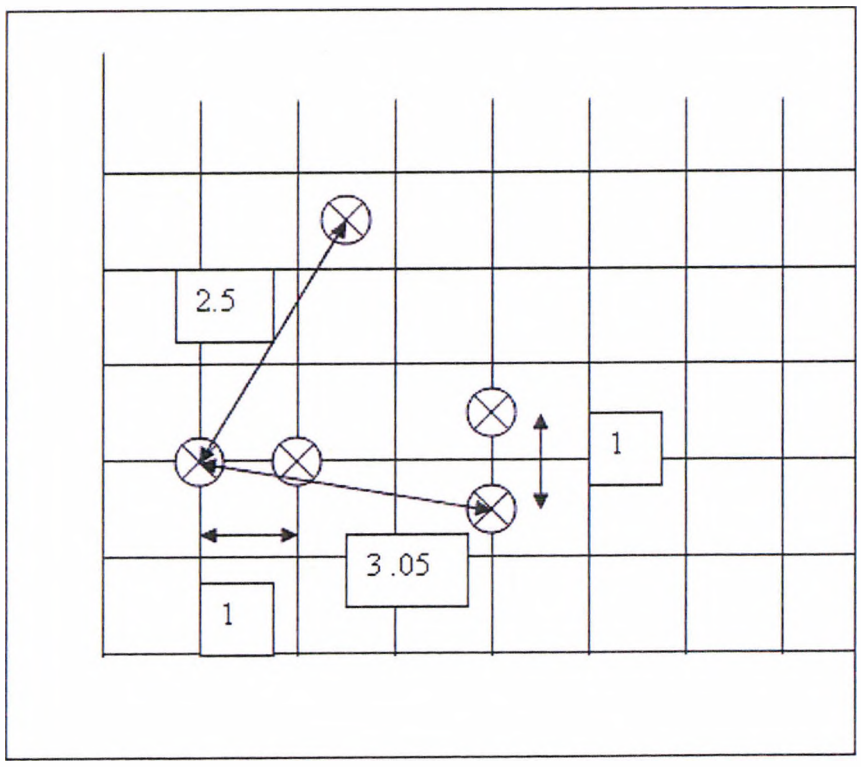


Figure A6.4 Distances between objects for 'complete' linking and Euclidean distance metric

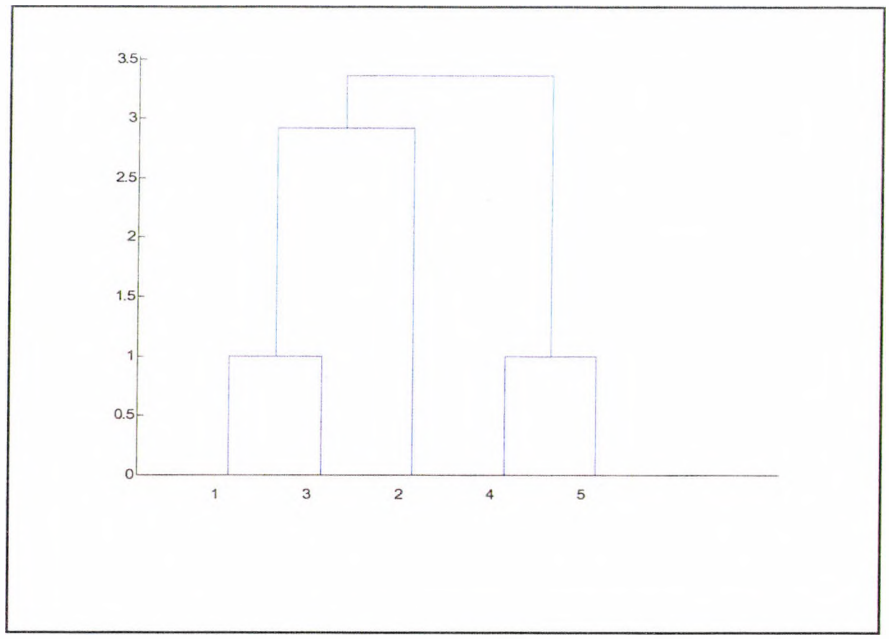


Figure A6.5 Dendrogram for 'complete' linking and Euclidean distance metric

Average Linkage

The 'average' linking gives: -

Object No.	Object No	Distance apart
1	3	1
4	5	1
6	2	2.5515
8	7	2.8298

Table A6.5 Distances of objects using 'average' linking

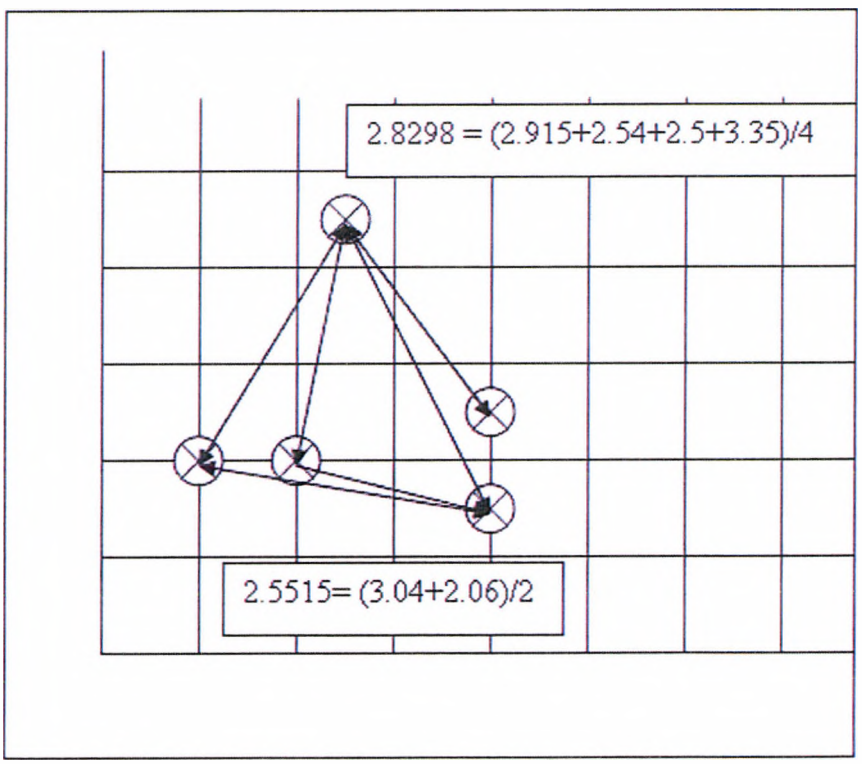


Figure A6.6 Distances between objects for 'average' linking and Euclidean distance metric

Average distance between objects 1 and 4 and 3 and 4 gives 2.5515and the average distance between object 2 and 1, 3, 4 and 5 is 2.8298.

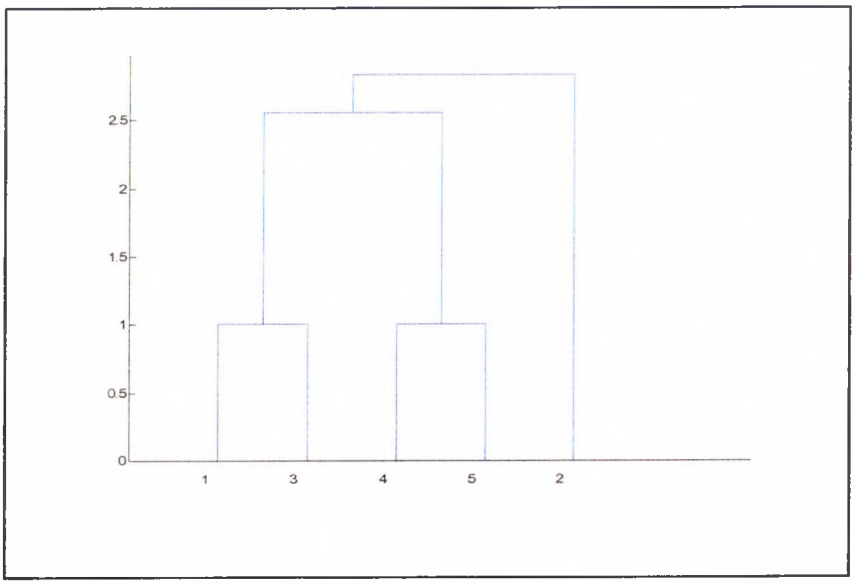


Figure A6.7 Distances between objects for 'average' linking and Euclidean distance metric

Centroid linking

The 'centroid' linking gives: -

Object No.	Object No	Distance apart
1	3	1
4	5	1
6	7	2.5
8	2	2.5125

Table A6.6 Distances of objects using 'centroid' linking and Euclidean distance metric

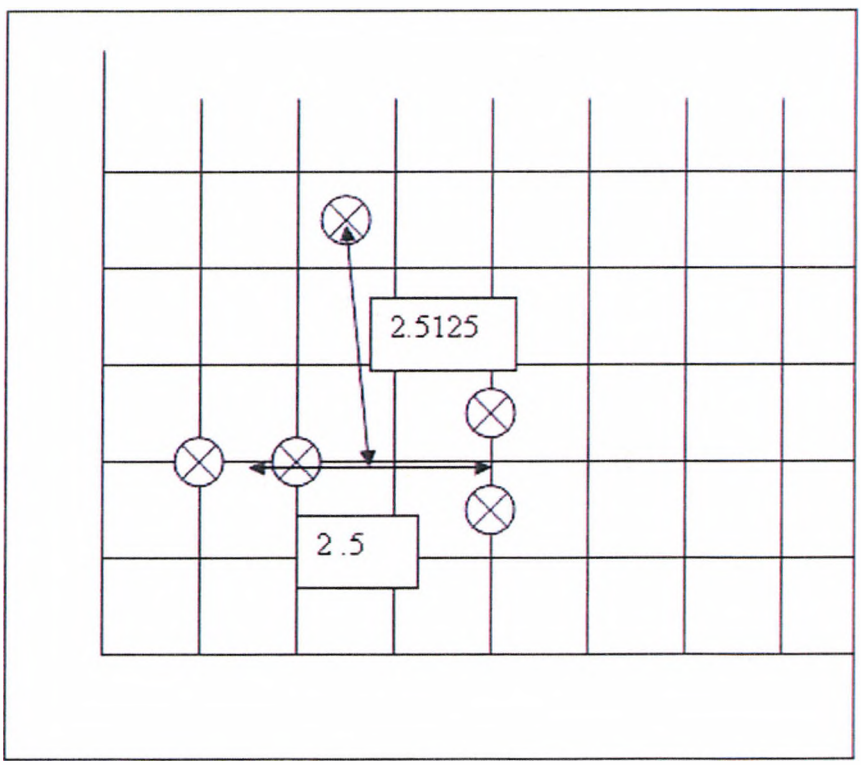


Figure A6.8 Distances between objects for 'centroid' linking and Euclidean distance metric

Centroid linkage and Euclidean distance:

The centroid of objects 1 and 3 is (1.5, 2) and the centroid between objects 4 and 5 is (4, 2). The centroid between the new object formed from objects 1, 3, 4 and 5 is (2.75, 2) and the distance from object 2 is 2.5125.

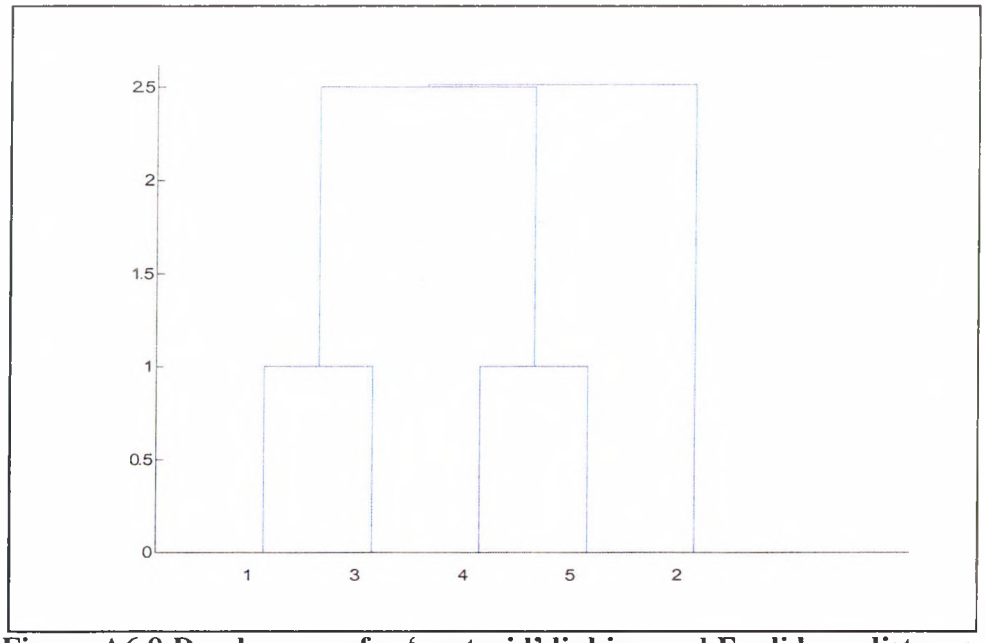


Figure A6.9 Dendrogram for 'centroid' linking and Euclidean distance metric

Ward or inner squared distance linking

The 'ward' linking gives: -

Object No.	Object No	Distance apart
1	3	0.707
4	5	0.707
6	2	2.1985
8	7	2.543

Table A6.7 Distances of objects using 'centroid' linking and Euclidean distance metric

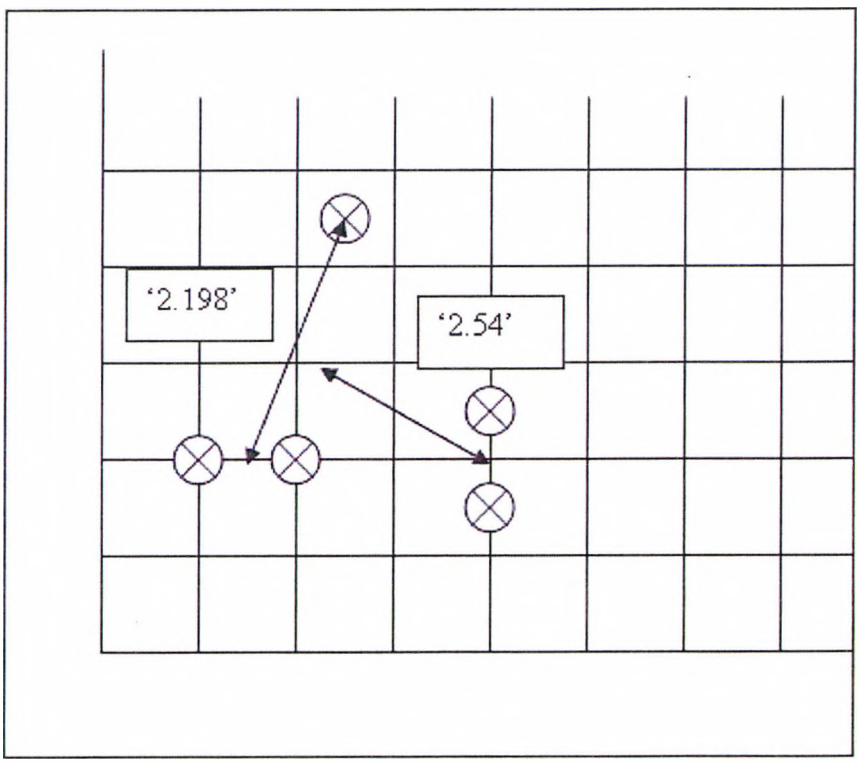


Figure A6.10 Distances between objects for 'ward' linking and Euclidean distance metric

Ward linkage and Euclidean distance:

The formula for 'ward' linkage is, $d_{rs}{}^2 = \frac{n_r n_s}{n_r + n_s}(dis\tan ce^2)$ which is the distance between cluster 'r' and cluster 's'.

Distance between 1 and 3 is $\sqrt{(1.1(1^2)/2)} = 0.707$

Distance between the centroid of 1 and 3 and object 2 is $\sqrt{(1.2(7.25)/3)} = 2.192$

Distance between the centroid of 1, 2 and 3 and centroid of objects 4 and 5 is $\sqrt{(3x2(5.389)/5)} = 2.192$

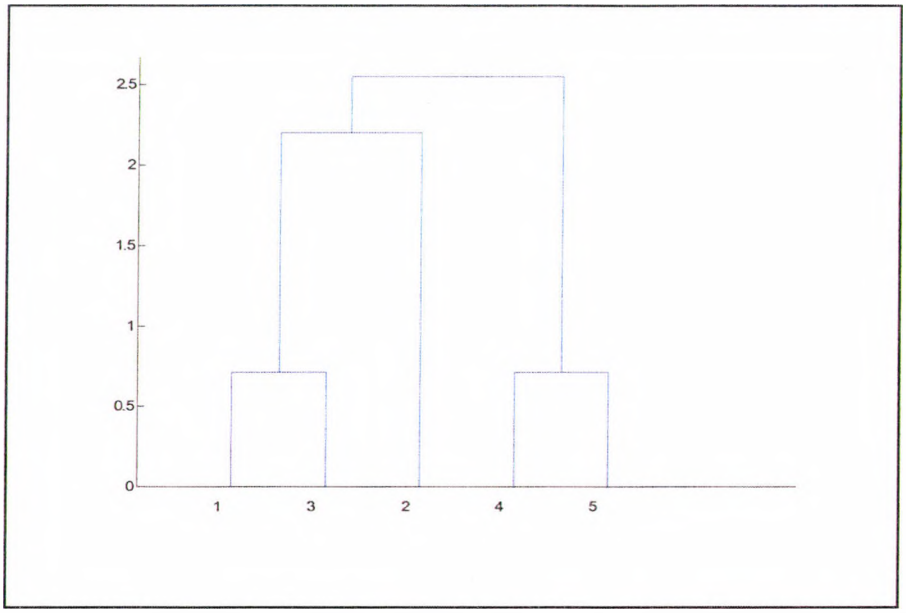


Figure A6.11 Dendrogram for 'ward' linking and Euclidean distance metric

Mahalanobis distance

This distance measure is only meaningful using single, complete or average linkage techniques.

The function 'pdist' returns distance information as a vector or matrix, Y.

0	2.5124	0.7695	2.3133	2.3821
2.5124	0	2.2042	2.7272	1.9845
0.7695	2.2042	0	1.5634	1.6308
2.3133	2.7272	1.5634	0	0.8557
2.3821	1.9845	1.6308	0.8557	0

Table A6.8 Mahalanobis distance matrix derived from the five objects

Single linkage

Object No.	Object No	Distance apart
1	3	0.7695
4	5	0.8557
6	7	1.5634
8	2	1.9845

Table A6.9 Distances of objects using 'single' linking with Mahalanobis distance metric

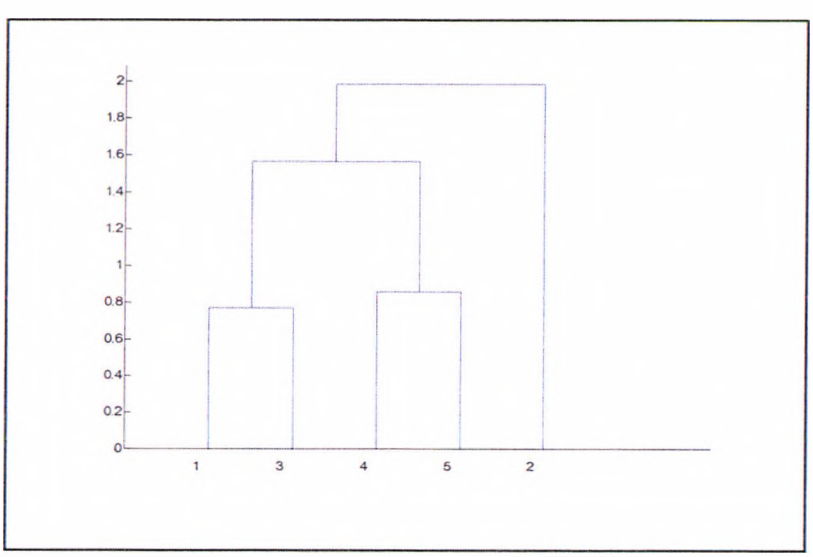


Figure A6.12 Dendrogram using 'single' linking with Mahalanobis distance metric

Complete linkage

Object No.	Object No	Distance apart
1	3	0.7695
4	5	0.8557
6	7	2.3821
8	2	2.7272

Table A6.10 Distances of objects using 'complete' linking and with Mahalanobis distance metric

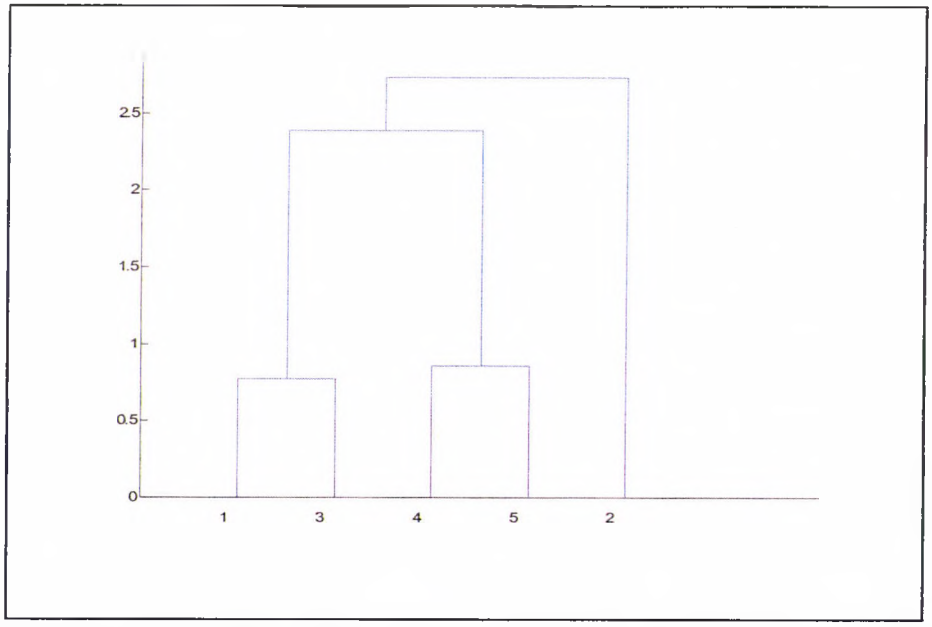


Figure A6.13 Dendrogram using 'complete' linking with Mahalanobis distance metric

Average linkage

Object No.	Object No	Distance apart
1	3	0.7695
4	5	0.8557
6	7	1.9724
8	2	2.3571

Table A6.11 Distances of objects using 'average' linking with Mahalanobis distance metric

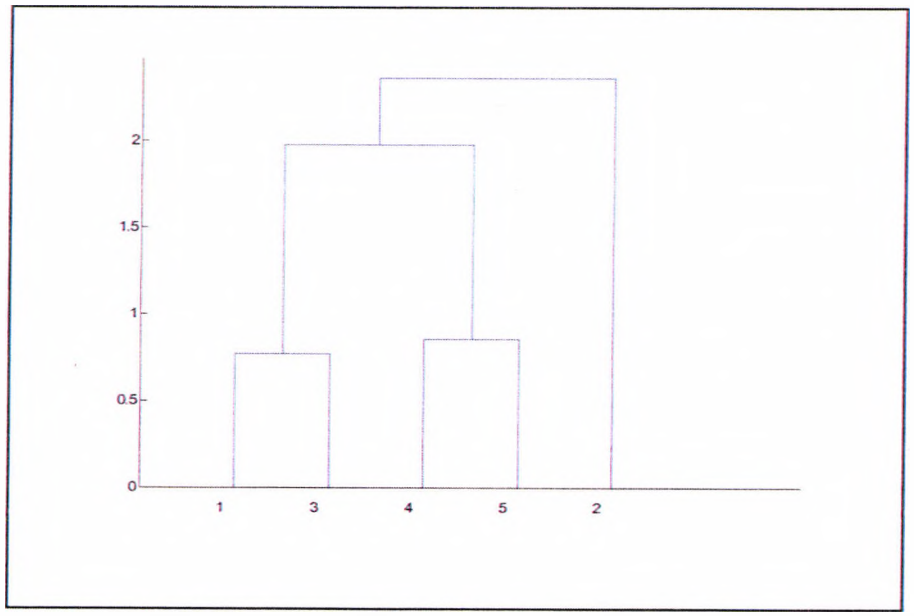


Figure A6.14 Dendrogram for 'average' linking with Mahalanobis distance metric

B Dendrogram changes due to linkage method and distance metric.

Table 6.6 'Comparison of distance metrics and linkage methods' in Chapter 6 was prepared from the following investigations. The linkage methods were changed for the three distance metrics of 'City Block', 'Euclidean' and 'Mahalanobis'.

1 City Block distance metric

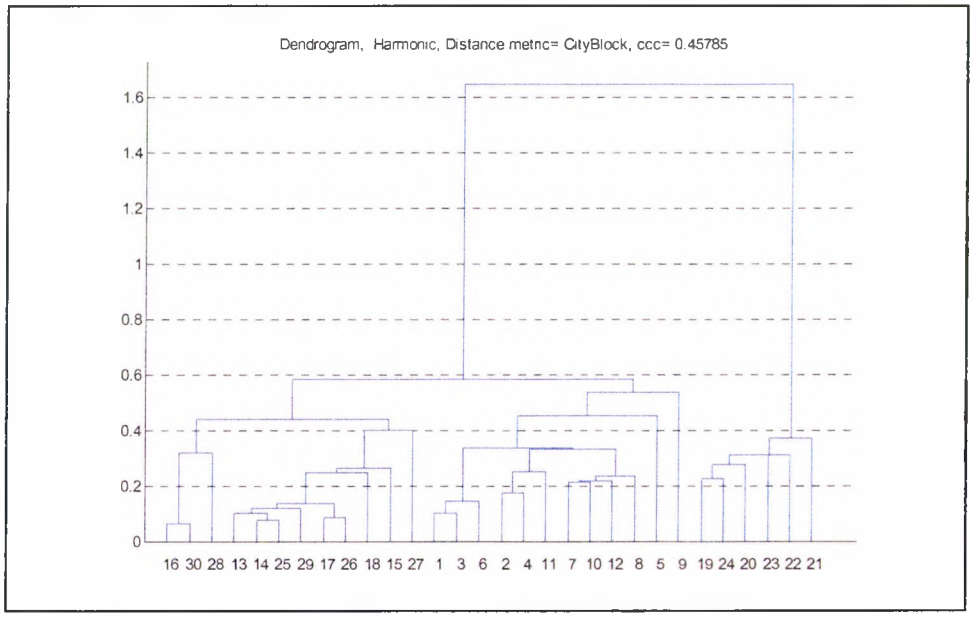


Figure A6.15 'City Block' distance metric, 'single' linkage method

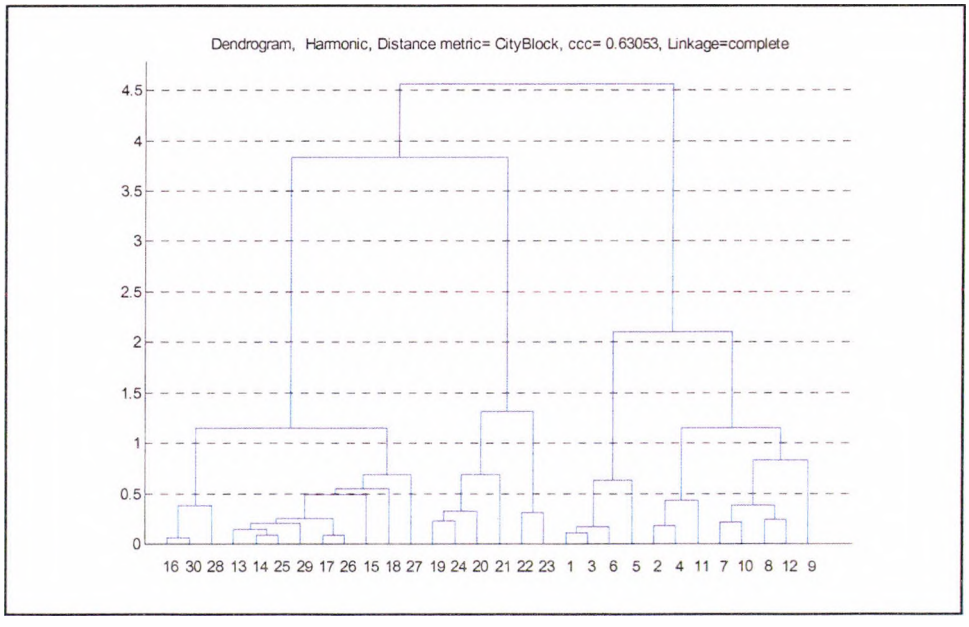


Figure A6.16 'City Block' distance metric, 'complete' linkage method

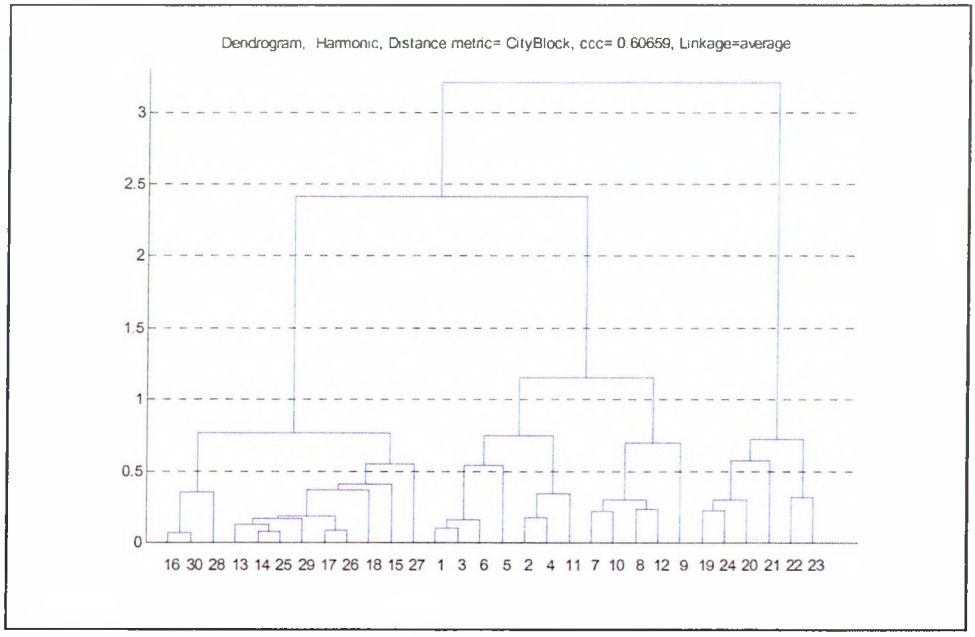


Figure A6.17 'City Block' distance metric, 'average' linkage method

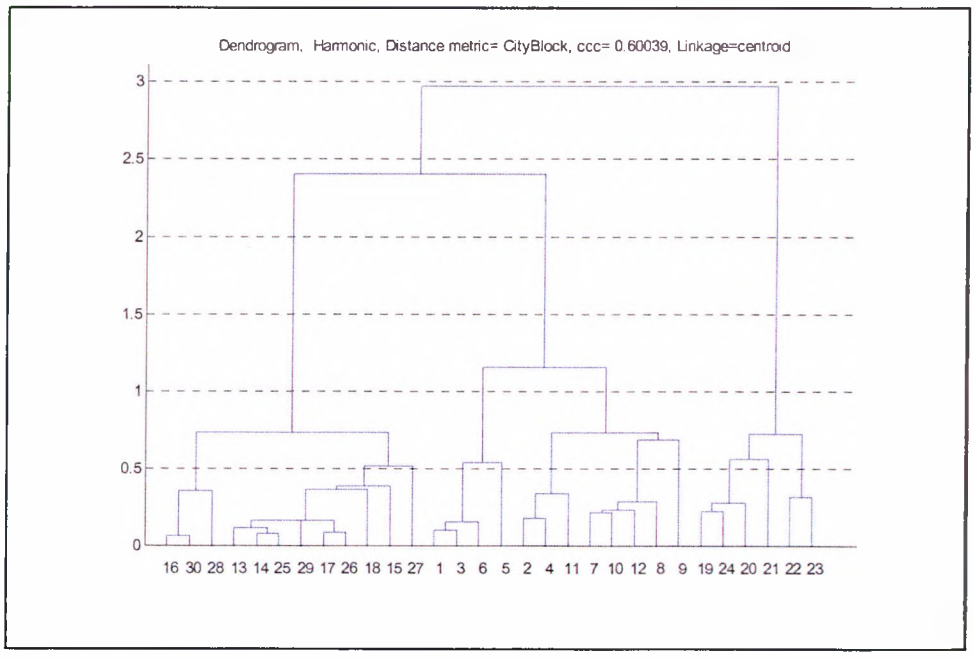


Figure A6.18 'City Block' distance metric, 'centroid' linkage method

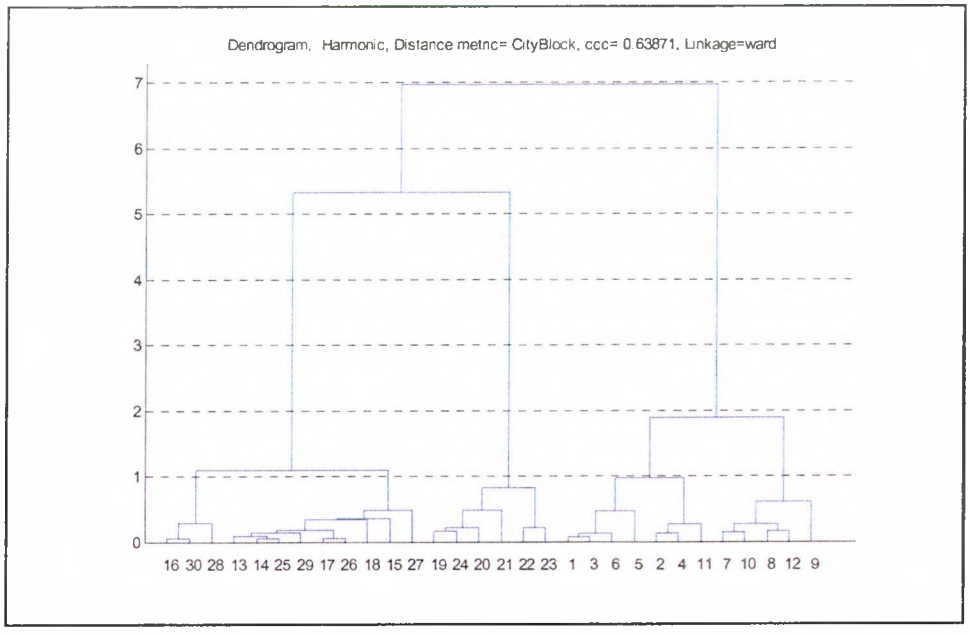


Figure A6.19 'City Block' distance metric, 'ward' linkage method

2 Euclidean distance metric

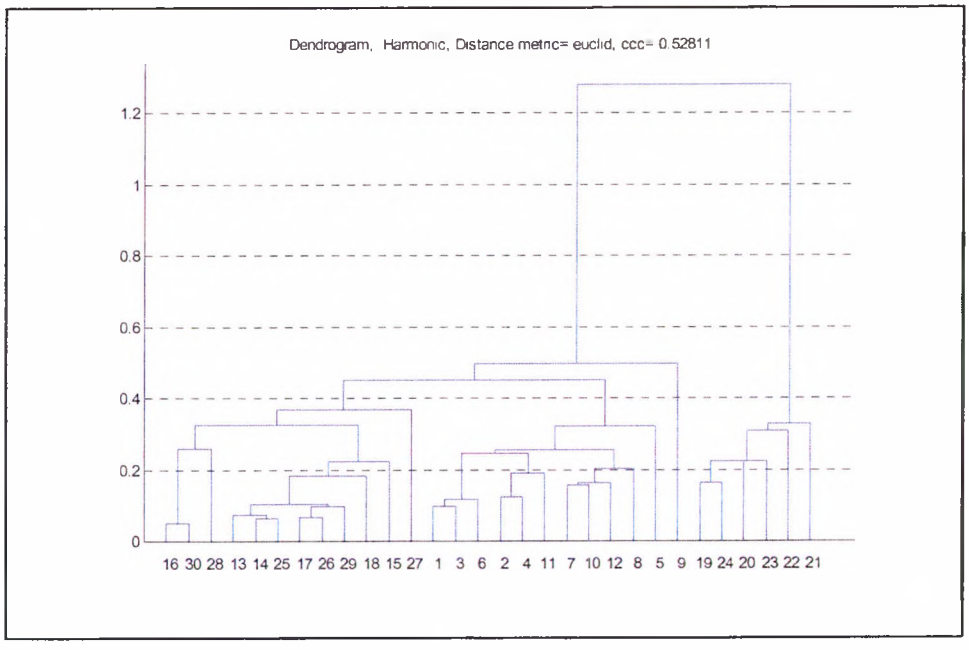


Figure A6.20 'Euclidean' distance metric, 'single' linkage method

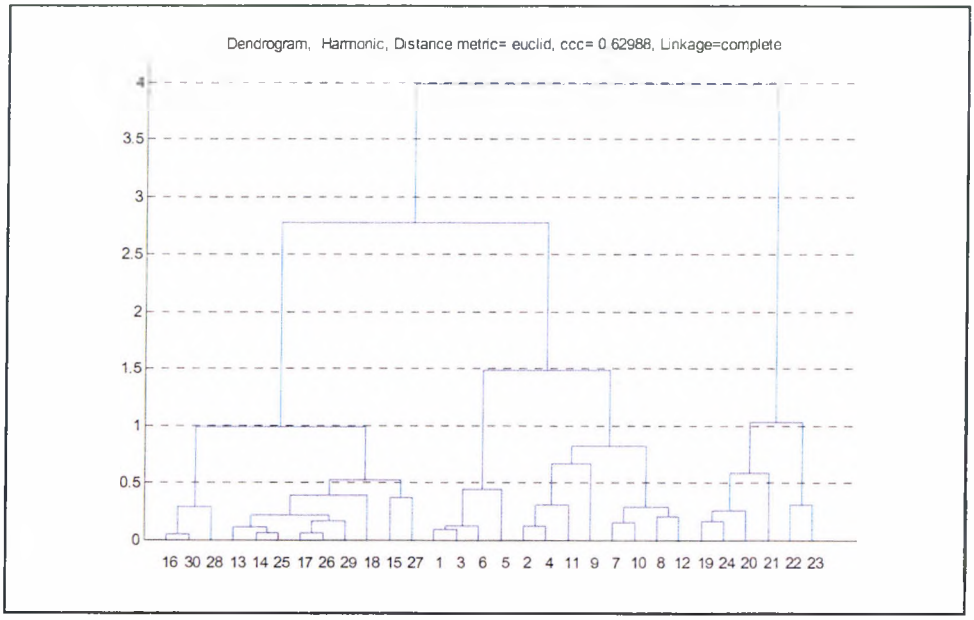


Figure A6.21 'Euclidean' distance metric, 'complete' linkage method

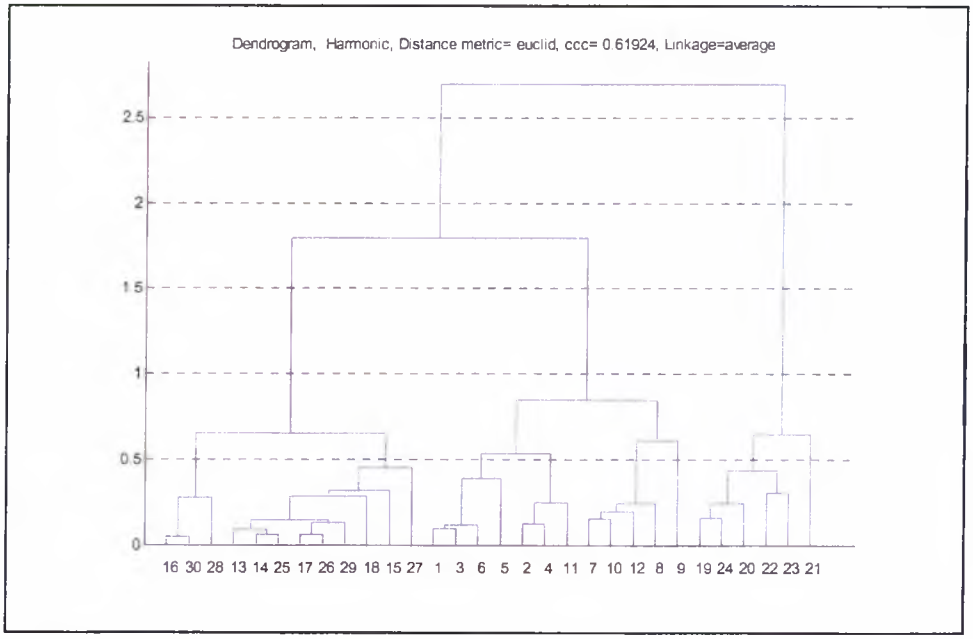


Figure A6.22 'Euclidean' distance metric, 'average' linkage method

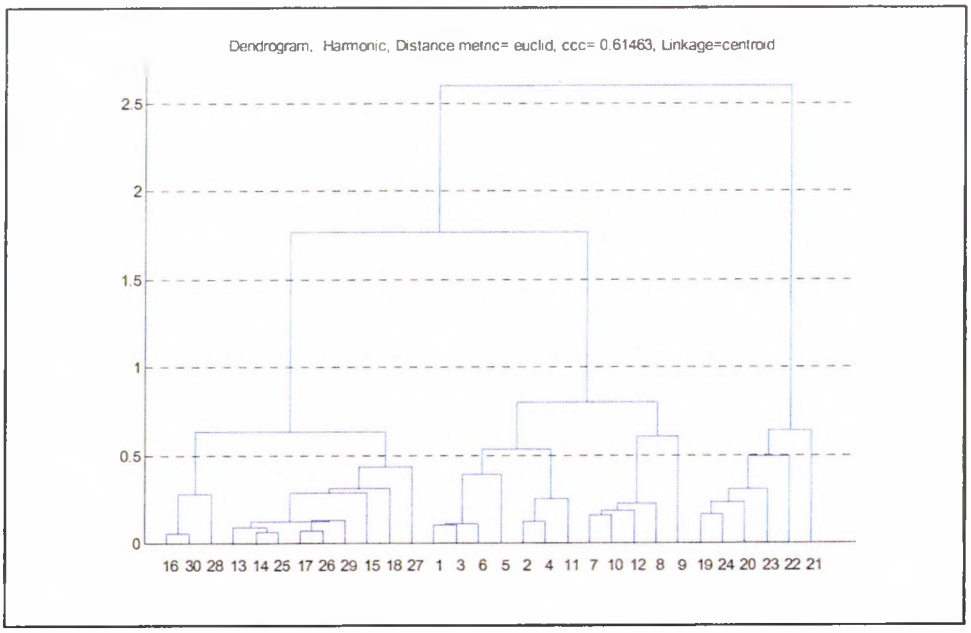


Figure A6.23 'Euclidean' distance metric, 'centroid' linkage method

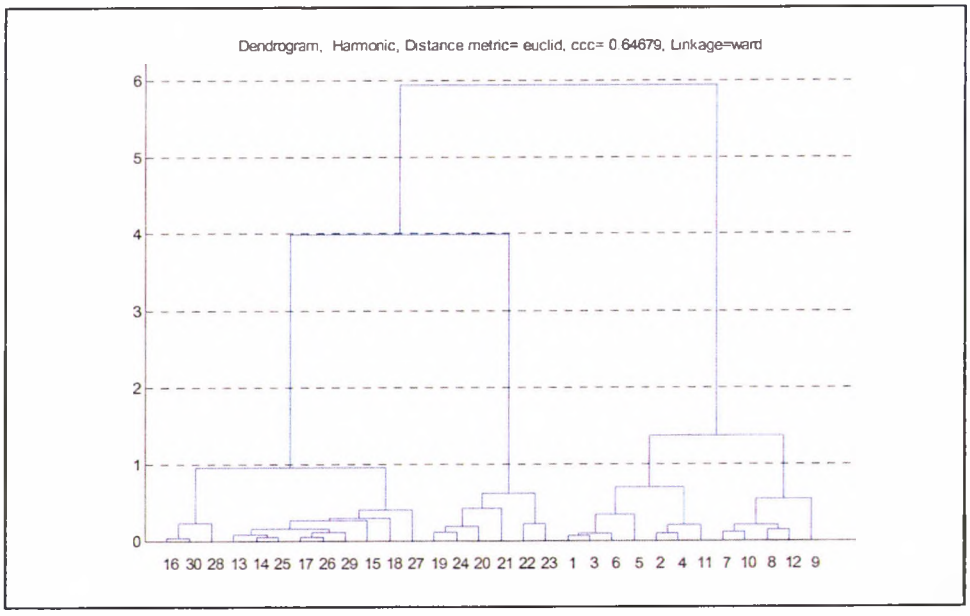


Figure A6.24 'Euclidean' distance metric, 'ward' linkage method

3 Mahalanobis distance metric

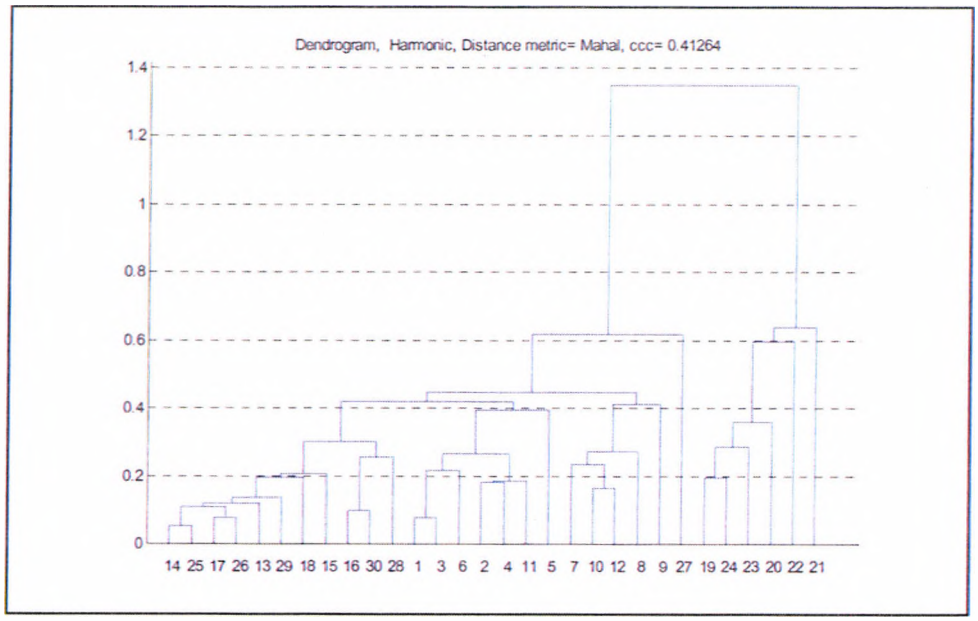


Figure A6.25 'Mahalanobis' distance metric, 'single' linkage method

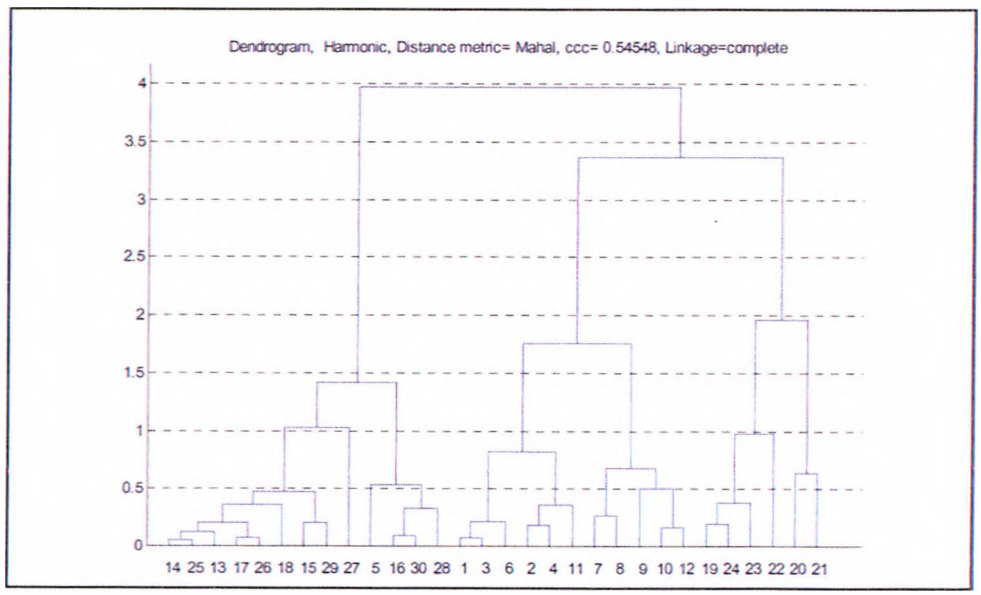


Figure A6.26 'Mahalanobis' distance metric, 'complete' linkage method

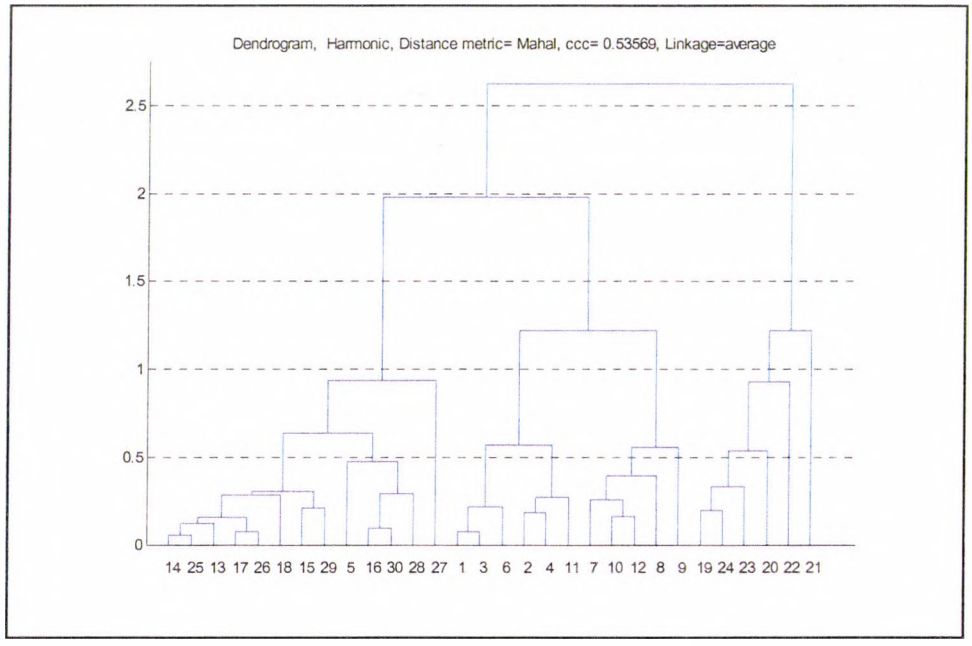


Figure A6.27 'Mahalanobis' distance metric, 'average' linkage method

Appendix VII – Four Gesture Experiments

1 Pointing Gesture Experiments

This experiment (Harding and Ellis, 2004) used 5 subjects who all undertook the same sequence of gestures. The subjects were seated on a chair and a web-cam used to record the five gestures that made up a gesture sequence, as shown in Figure A7.1.

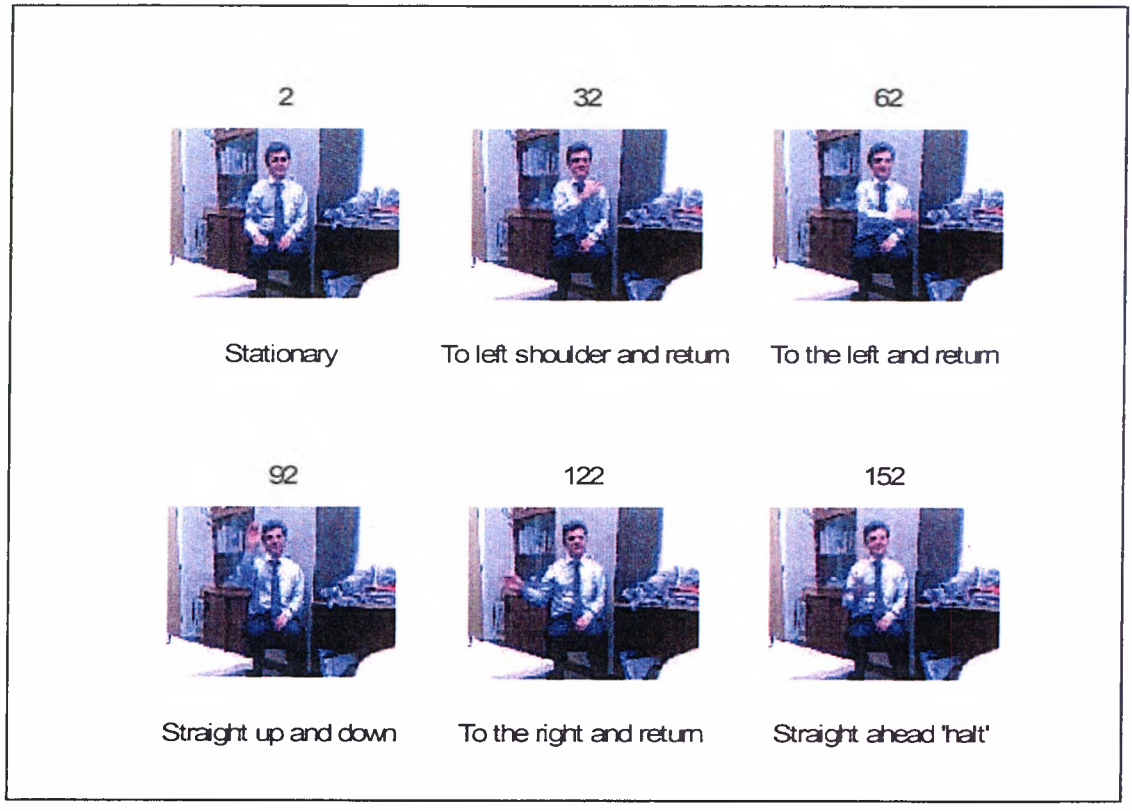


Figure A7.1 Illustration of a gesturer stationary (frame 2) and in the process of enacting the five gestures (frames 32, 62, 92, 122 and 152)

Two of the subjects produced a sequence of gestures in about 160 frames. These gestures were labelled A and B. Gesturer B was also recorded at a different time, and at a different rate, with three other gesturers and labelled as the gestures C, D, E and F respectively. The sequence of the six gestures for each gesturer was typically 360 frames. The complete gesture sequences were segmented into the individual gestures 1, 2, 3, 4 and 5 for each gesturer A, B, C, D, E and F.

The row and column coordinates are shown in figure A7.2.

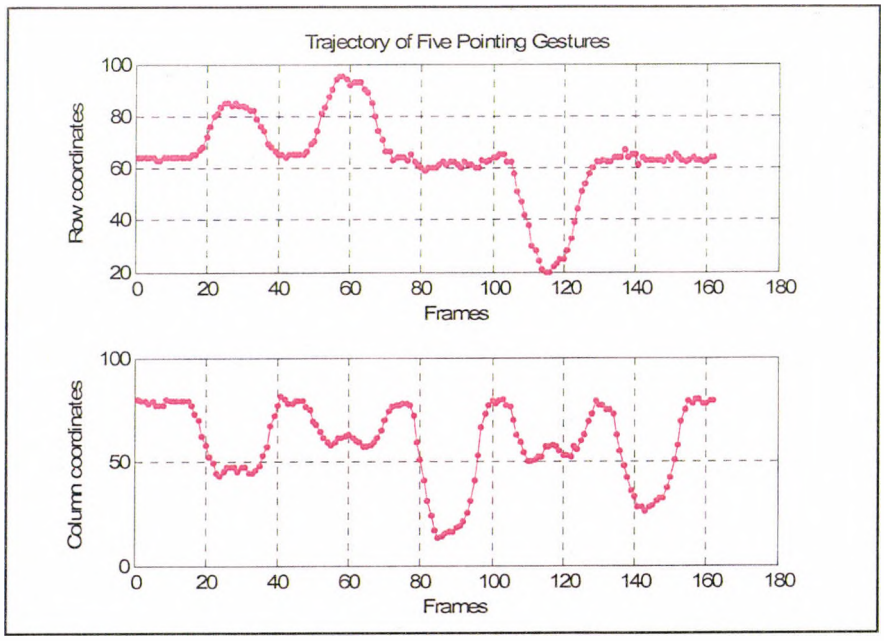


Figure A7.2 The row and column coordinates of the continuous trajectory of five pointing gestures for one gesturer.

A complete sequence of images for one of the gestures is shown in Figures A7.3 to Figure A7.7.

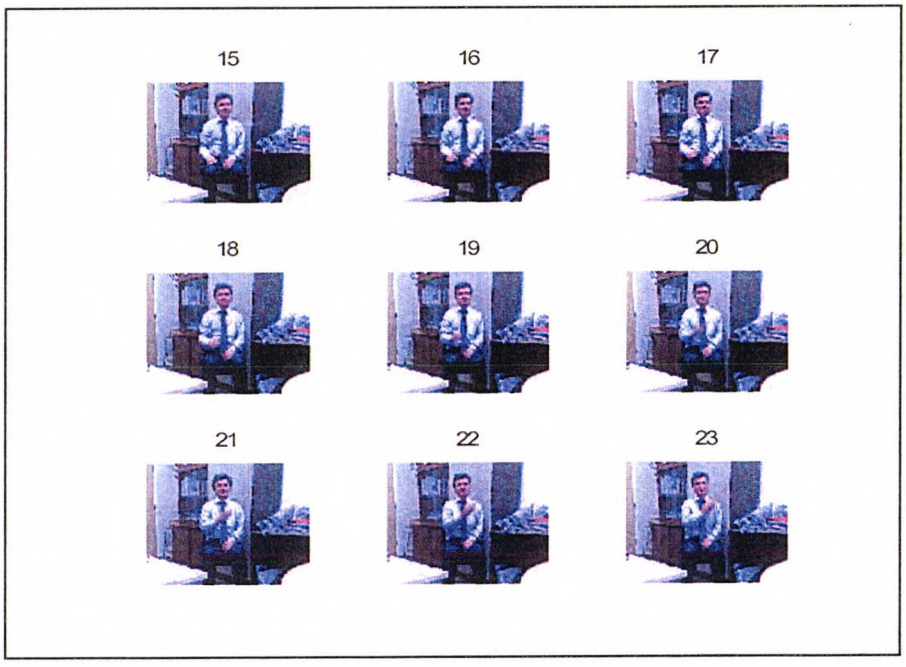


Figure A7.3 Frames 15 to 23 of gesture type 1

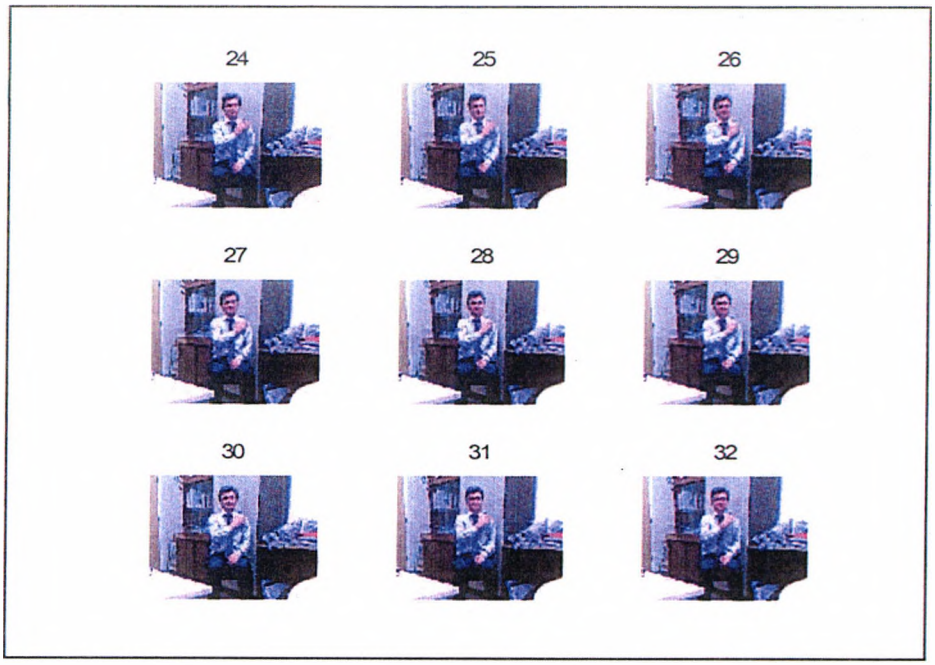


Figure A7.4 Frames 24 to 32 of gesture type 1

Figure A7.5 Frames 33 to 41 of gesture type 1

2 Ten repeated hand raising gesture experiments

The ten repeated hand-raising gestures were obtained from: -

ftp://pets.rdg.ac.uk/PETS-ICVS [Accessed **October 2002**]

X:\images\petsicvs\data\ScenarioA1\Cam1 for frames 16381 to 17240

The description of the various scenarios is as follows. However, the scenario A does not appear to relate to the actual image sequence that has been used. It would appear that scenario B would give a better description of what was in the images.

Scenario A: "Performing distinct Facial Expressions"

Actions: Sitting down, getting up, smile, angry, neutral, looking at other participants

Each person (1 to 6) enters in the room one after each other, go to his place,

presents himself to the frontal camera, and sit down.

Then each person looks at each person in front of him with a different facial

expression:

Scenario B: "Performing face & hand gestures"

Actions: Sitting down, getting up, raising hand, shaking head, nodding head, yawning, laughing

Estimated duration: 540 seconds, 1.9 Gb (AVI), 700 Mb (JPEG)

Real duration/size: 333 seconds, 1.0 Gb (JPEG)

Each person (1 to 6) enters in the room one after each other, go to his place,

presents himself to the frontal camera, and sit down.

For each person (1 2 3 4 5 6)

$person is raising left hand

$person is raising right hand

$person is shaking head

$person is nodding head

$person is yawning

$person is laughing

Frequency components

Typical frequency components for the fifth hand raising gesture are shown in Figure A7.6.

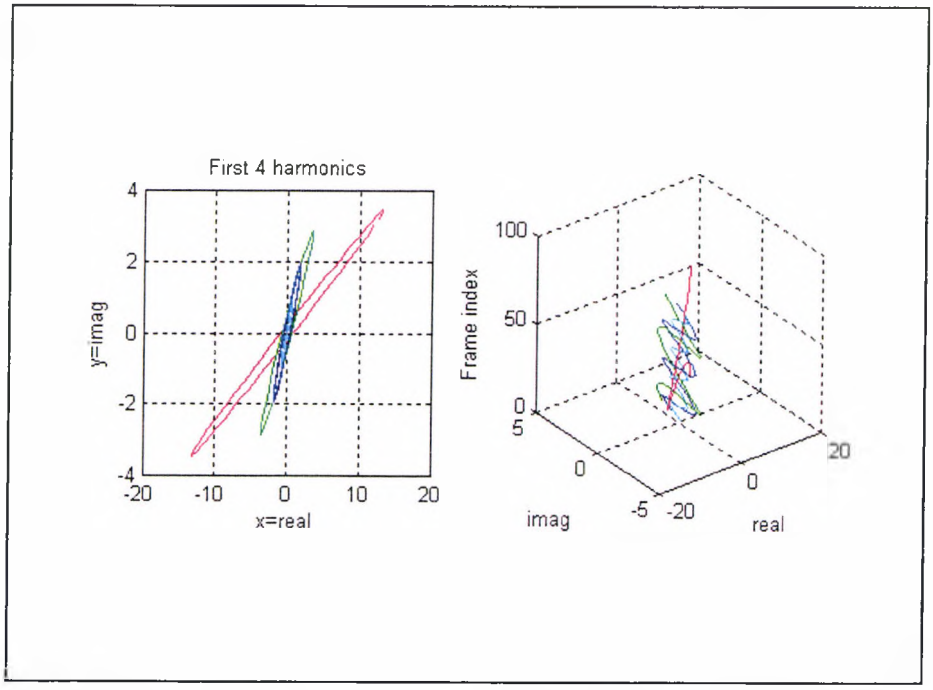


Figure A7.6 2D and 3D representation of the first three harmonic components of the fifth PETS hand raising gesture.

The first six orientation angles and associated magnitude of these components are shown in Table A7.1 and Table A7.2, for data that was obtained visually. The associated data that was obtained by automatic means using 'SCM' objects is shown in Tables A7.3 and A7.4.

Gesture	θ1	θ2	θ3	θ4	θ5	θ6
1	-21	-22	8	-48	24	21
2	-17	-12	-77	253	-58	-24
3	-19	-10	32	156	-200	-5
4	-15	-4	-12	-24	13	66
5	-20	-30	-118	294	-101	-33
6	-11	-9	-46	77	-38	-66
7	-6	-11	-21	-69	-33	123
8	-15	-4	-12	-24	13	66
9	-6	-11	-51	-70	-18	25
10	-11	-24	73	-63	-69	50
Average	-14	-14	-22	48	-46	22
Std	5	9	54	139	66	56

Table A7.2 Orientation angles (θ) for 6 harmonics and 10 gestures, data obtained visually.

Gesture	M1	M2	M3	M4	M5	M6
1	1.57	0.60	0.28	0.09	0.06	0.05
2	1.30	0.49	0.06	0.06	0.04	0.03
3	1.31	0.38	0.09	0.02	0.02	0.04
4	1.33	0.88	0.38	0.14	0.05	0.06
5	1.36	0.33	0.15	0.06	0.03	0.01
6	1.42	0.67	0.15	0.11	0.05	0.03
7	1.40	0.71	0.23	0.05	0.01	0.02
8	1.33	0.88	0.38	0.14	0.05	0.06
9	1.35	0.65	0.18	0.08	0.05	0.03
10	1.34	0.43	0.18	0.07	0.05	0.03
Average	1.37	0.60	0.21	0.08	0.04	0.04
Std	0.08	0.20	0.11	0.04	0.01	0.02

Table A7.3 Magnitude, (M = positive + negative sequence value) for 6 harmonics for 10 gestures, data obtained visually. Normalised by A_p =1.

Gesture	θ1	θ2	θ3	θ4	θ5	θ6
1	-15	-27	6	-24	90	-63
2	-15	-10	9	-53	42	65
3	-15	-24	26	6	8	-33
4	-16	-5	-20	-62	88	-2
5	-15	-24	-8	-5	-114	147
6	-6	2	-29	-40	72	38
7	-5	-12	-20	-42	68	-30
8	-7	-9	-31	59	-35	-12
9	0	-12	-29	-10	10	21
10	-1	-11	-29	86	-54	5
Average	-9	-13	-12	-9	17	14
Std	6	9	19	48	68	59

Table A7.4 Orientation angles (θ) for 6 harmonics for 10 gestures, data obtained automatically.

Gesture	M1	M2	M3	M4	M5	M6
1	1.59	0.59	0.34	0.13	0.05	0.13
2	1.34	0.42	0.13	0.08	0.12	0.07
3	1.39	0.36	0.19	0.10	0.08	0.09
4	1.34	0.79	0.35	0.12	0.14	0.09
5	1.39	0.46	0.27	0.12	0.10	0.15
6	1.37	0.72	0.24	0.10	0.10	0.09
7	1.39	0.52	0.12	0.04	0.07	0.08
8	1.49	0.68	0.28	0.14	0.13	0.11
9	1.37	0.63	0.18	0.12	0.05	0.08
10	1.29	0.57	0.21	0.07	0.15	0.15
Average	1.39	0.57	0.23	0.10	0.10	0.11
Std	0.09	0.14	0.08	0.03	0.04	0.03

Table A7.5 Magnitude (M = positive + negative sequence value) for 6 harmonics for 10 gestures, data obtained automatically. Normalised by A_p =1

The distribution of first, second and third harmonic vectors for the automatically generated harmonic components are shown in Figure A7.7

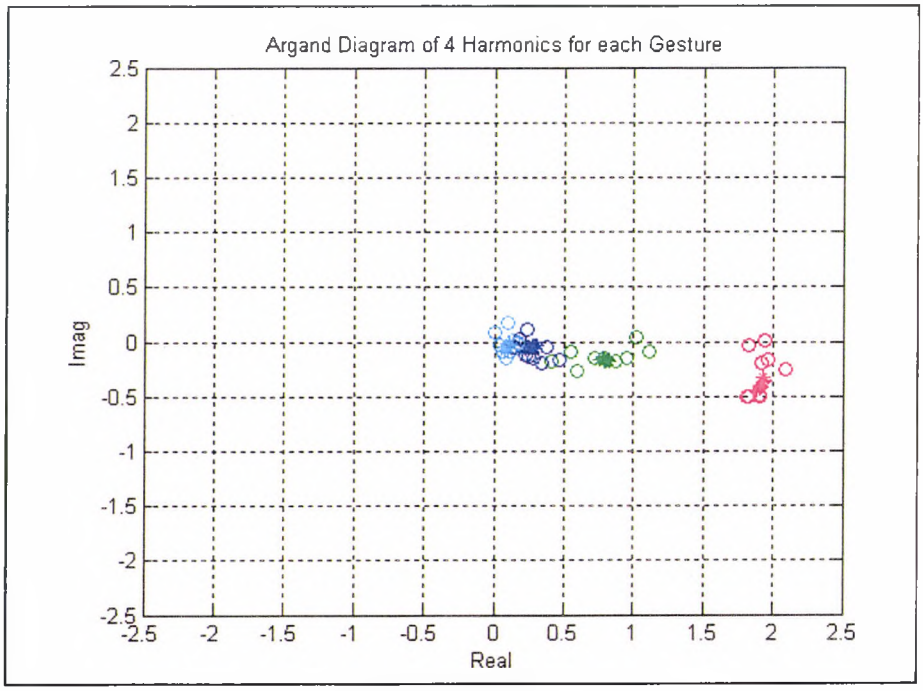


Figure A7.7 Distribution of the first three harmonic vectors for the ten hand raising gestures.

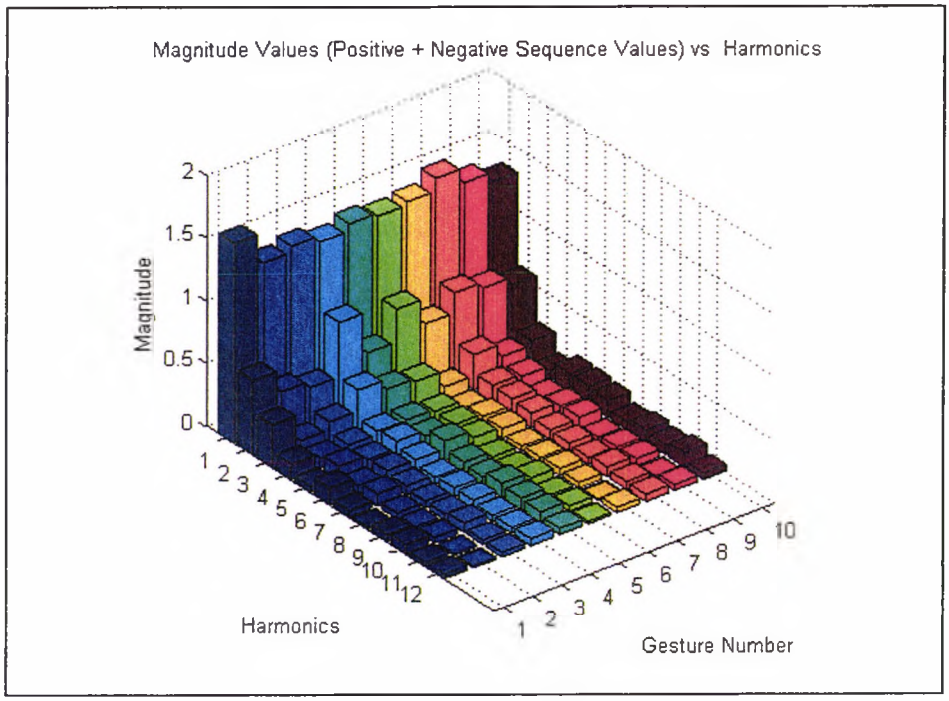


Figure A7.8 Distribution of the vector magnitudes for the ten hand raising gestures.

3 'Take Mug' Sequence Experiments

A montage of images from Figures A7.9 to A7.14 showing the 'Take Mug' gesture.

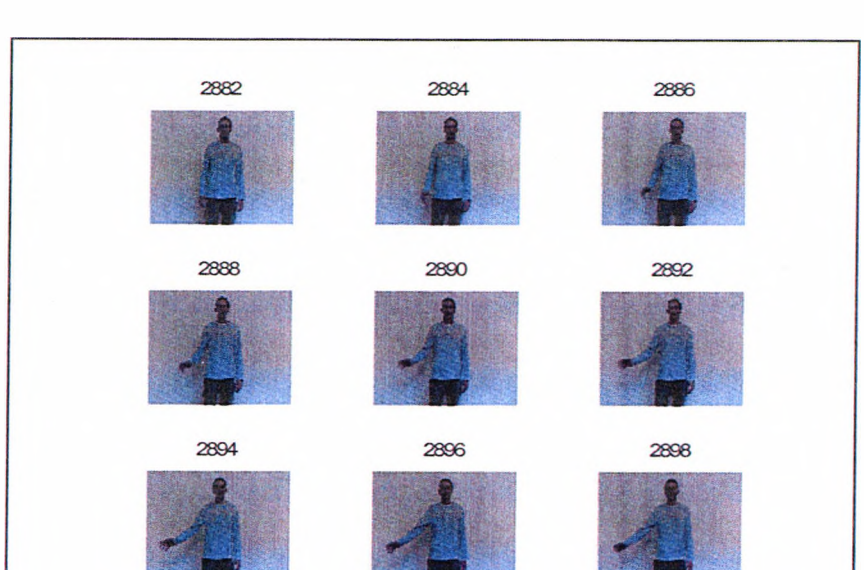


Figure A7.9 Alternate frames 2882 to 2898 from a 'Take-Mug' gesturer

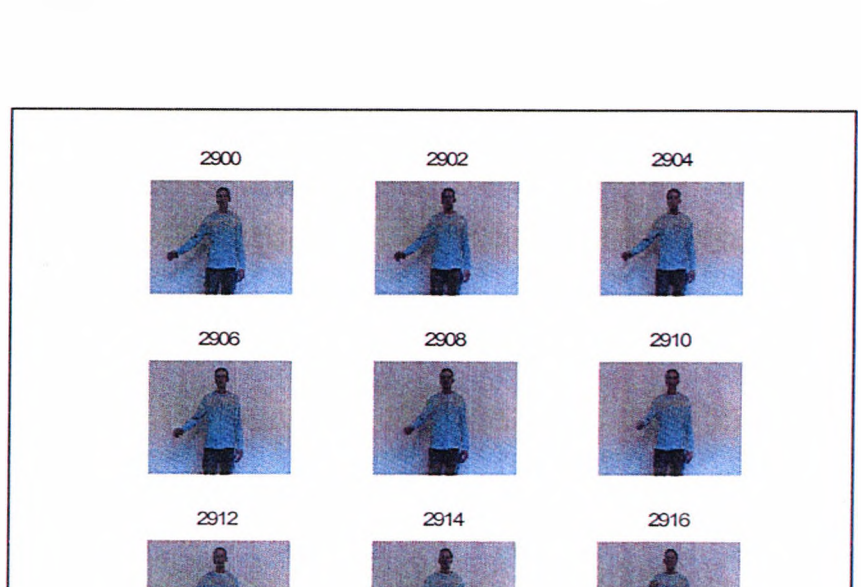


Figure A7.10 Alternate frames 2900 to 2916 from a 'Take-Mug' gesturer

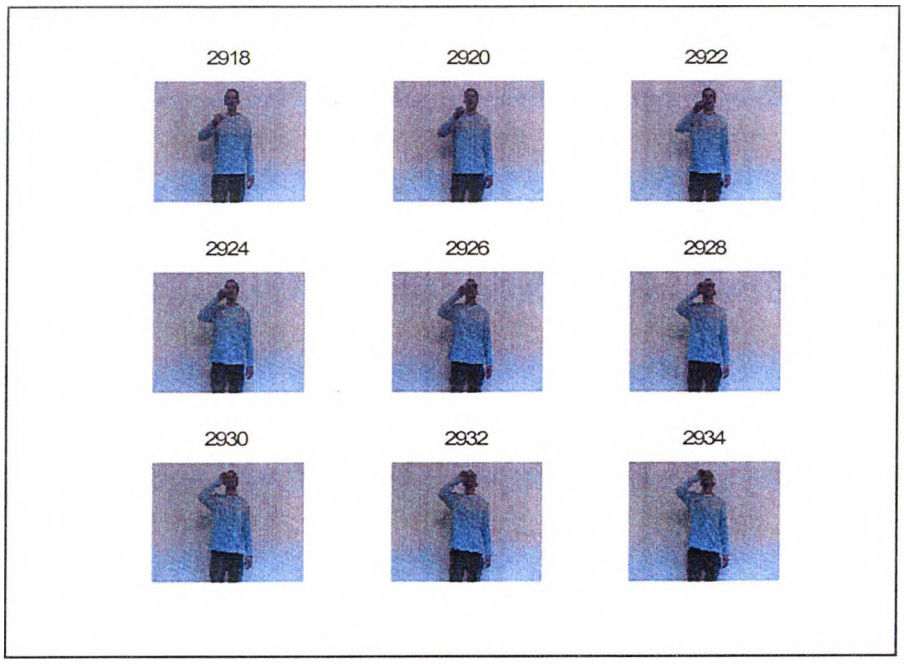


Figure A7.11 Alternate frames 2918 to 2934 from a 'Take-Mug' gesturer

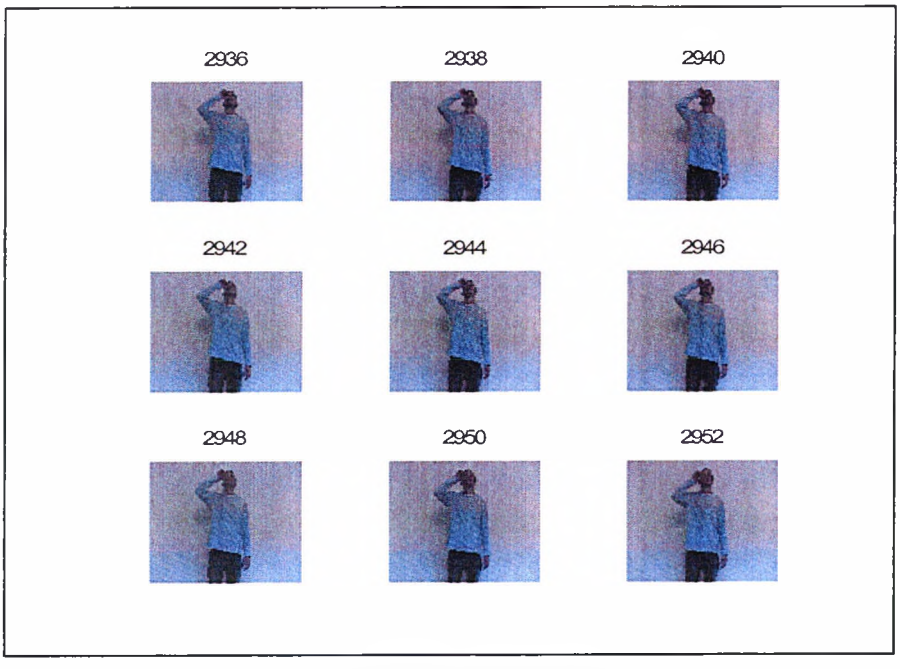


Figure A7.12 Alternate frames 2936 to 2952 from a 'Take-Mug' gesturer

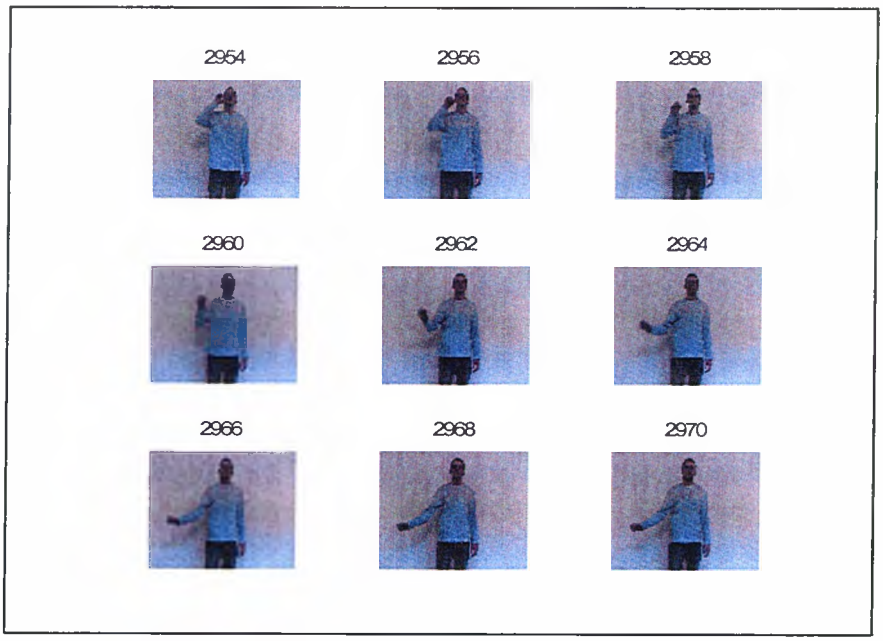


Figure A7.13 Alternate frames 2954 to 2970 from a 'Take-Mug' gesturer

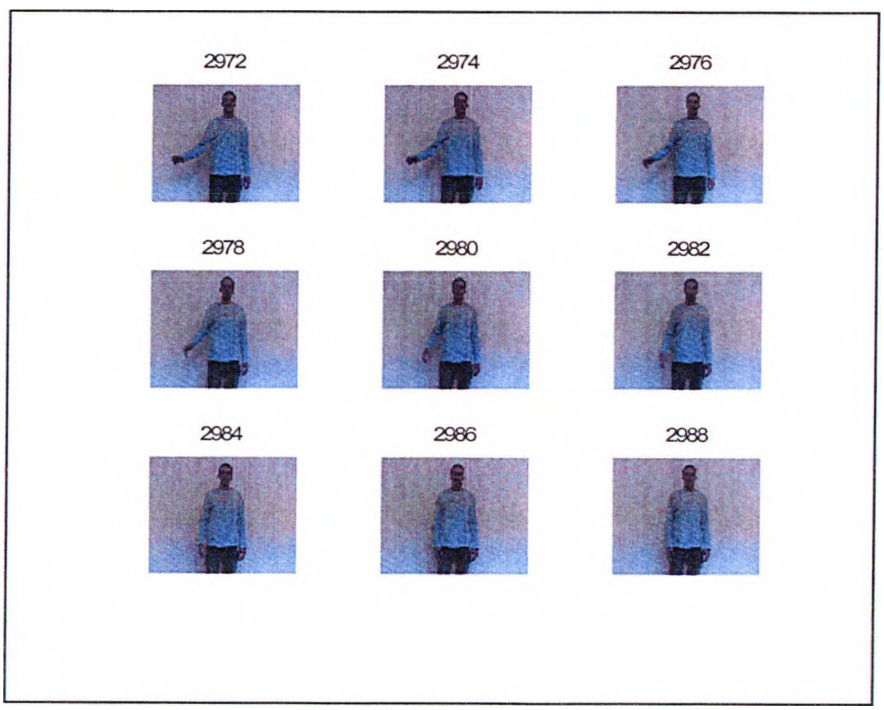


Figure A7.14 Alternate frames 2972 to 2988 from a 'Take-Mug' gesturer

Visually Observed Characteristics of 'Take Mug' Suite of Experiments

Person	Comment and characteristics
A	Take Mug steady up and down
B	Similar to 1, but curved at end
C	Take Mug, steady but wobble up and down
D	Inflection going up; replaced Mug going down
E	Take Mug, curved at end
F	Inflection going up and fast. Similar going down but did not stop.
G	Deliberate Take Mug up and replaced going down
H	Deliberate take and replace. Action in front.
J	Deliberate take and replace (longer).
K	Kink at beginning. Take and replace, different spatial positions.
L	Quick beginning. Short stay at top, different spatial positions.
M	Very wide, up and down very similar.
N	Truncated at beginning, but regular.
O	Kink at beginning, Take and replace.
P	Fast at beginning, different spatial positions, outward kink at end
Q	Take, very slow to top, straight down (no replacement)
R	Very wide, after replacement long stop.
S	Kink to take, no replacement straight down
T	Take, head back, replace
U	Wide, take, head back, replace.
V	Wide, take; head back, replacement short hesitation.

Table A7.6 Description of the characteristics and differences of each individual 'Take Mug' gesturer.

Statistics of the orientation angles and associated magnitudes are shown in Table A7.6, supported by the distribution of the vectors in figure A7.6

Harmonic	Magnitude Average	Magnitude Std.	Orientation Angle, average	Orientation Angle, std
1	1.4184	0.093	21.3152	6.4396
2	0.3408	0.2058	-46.6099	84.0808
3	0.2439	0.1161	8.8246	61.466
4	0.1747	0.0722	15.4327	74.5303
5	0.0968	0.0621	-22.5211	49.4247
6	0.0628	0.0284	-6.9011	64.7081
7	0.0425	0.0249	19.0562	45.9572
8	0.0297	0.0163	10.2157	63.9804
9	0.023	0.0101	-26.1395	90.9237
10	0.021	0.0086	12.0513	88.0041
11	0.0193	0.0079	-24.3931	50.6091
12	0.016	0.0069	4.0513	46.6973

Table A7.7 Average and Standard Deviation (std.) of twelve harmonics from the twenty one 'Take Mug' gestures.

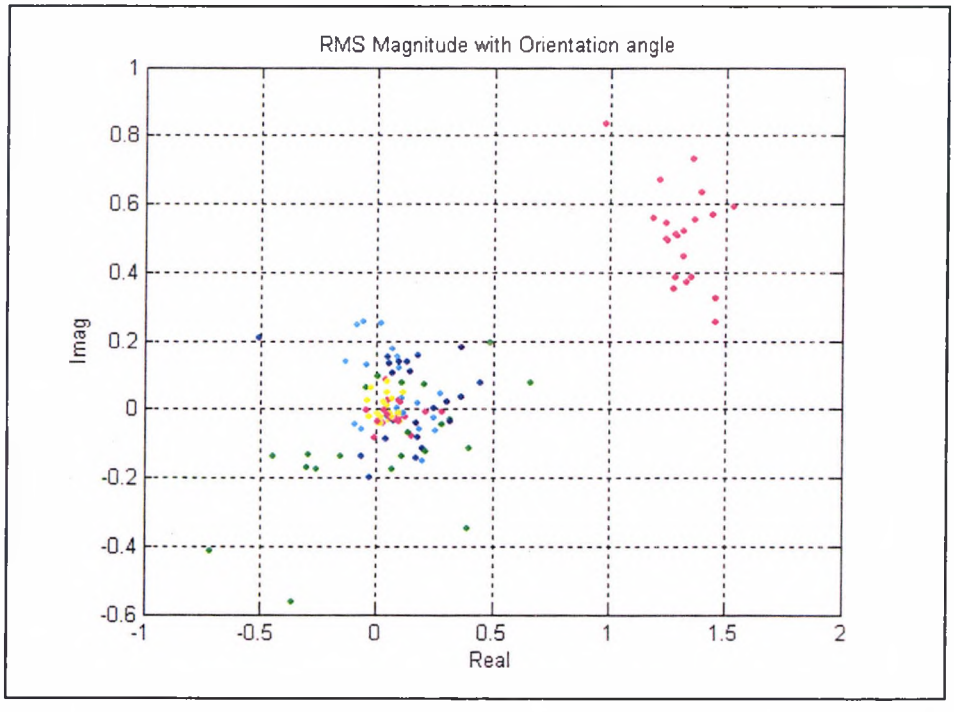


Figure A7.15 Distribution of the first six harmonics of the twenty-one 'Take Mug' gestures. Normalised by A_p =1

Visually determined sub-groupings of 'Take Mug' suite.

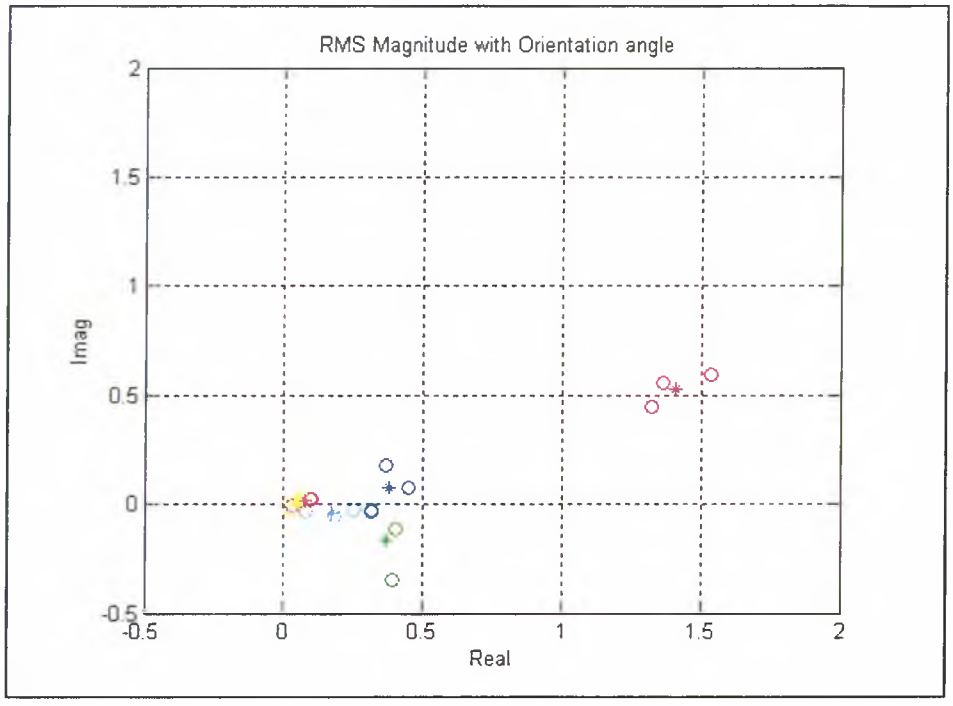


Figure A7.16 Visual grouping of gestures similar to gesture A.

Harmonic	Magnitude Average	Magnitude Std.	Orientation Angle, average	Orientation Angle, std
1	1.504	0.1278	20.5603	1.8154
2	0.4188	0.104	-21.0827	18.7238
3	0.3938	0.0692	9.7461	16.3295
4	0.1769	0.0851	-16.2121	9.2357
5	0.0796	0.0398	6.7083	11.3459
6	0.0525	0.0183	-7.5994	44.2406
7	0.0406	0.014	31.2426	52.6719
8	0.0296	0.006	-5.603	29.9713
9	0.0228	0.0087	-36.6593	112.9821
10	0.0242	0.0093	30.1205	140.6033
11	0.0261	0.0029	-2.1379	60.2201
12	0.0137	0.0042	-5.7272	34.6703

Table A7.8 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture A. Normalised by A_p =1

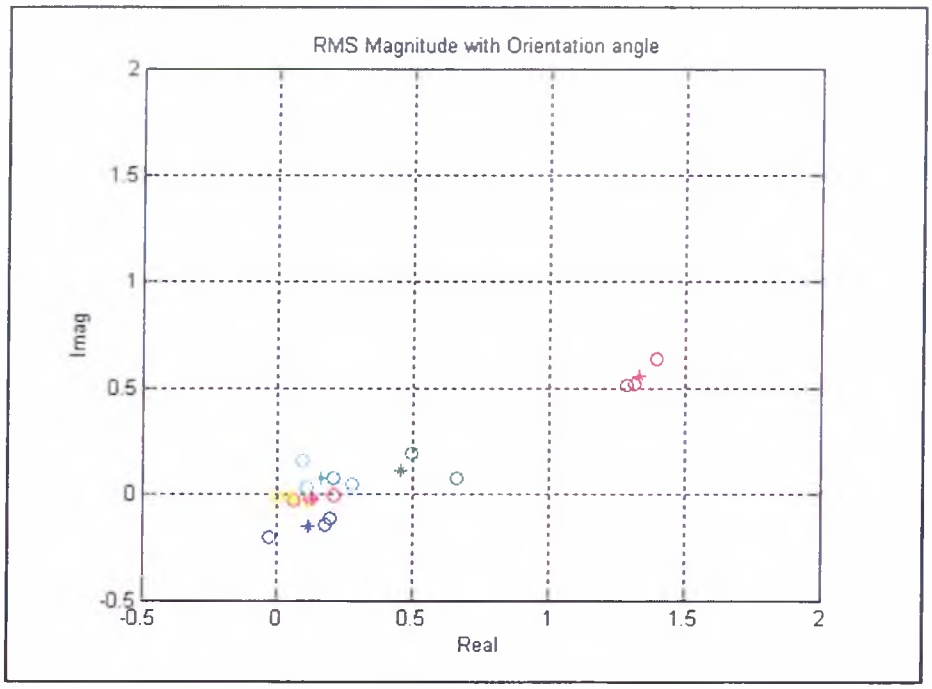


Figure A7.17 Visual grouping of gestures similar to gesture G.

Harmonic	Magnitude Average	Magnitude Std.	Orientation Angle, average	Orientation Angle, std
1	1.4457	0.0791	22.5429	1.6088
2	0.4697	0.2311	15.6479	8.0386
3	0.2199	0.0145	-56.5325	37.1052
4	0.1896	0.0846	27.7645	27.5107
5	0.135	0.0731	-12.0346	10.3064
6	0.0612	0.0473	-36.711	34.1207
7	0.0424	0.018	-0.1576	60.8885
8	0.0312	0.0144	26.1245	87.787
9	0.0285	0.0031	-9.7557	150.7826
10	0.02	0.0094	-36.1973	75.3214
11	0.0194	0.0078	17.1326	35.8536
12	0.0199	0.0078	-6.8551	32.8446

Table A7.9 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture G. Normalised by A_p =1

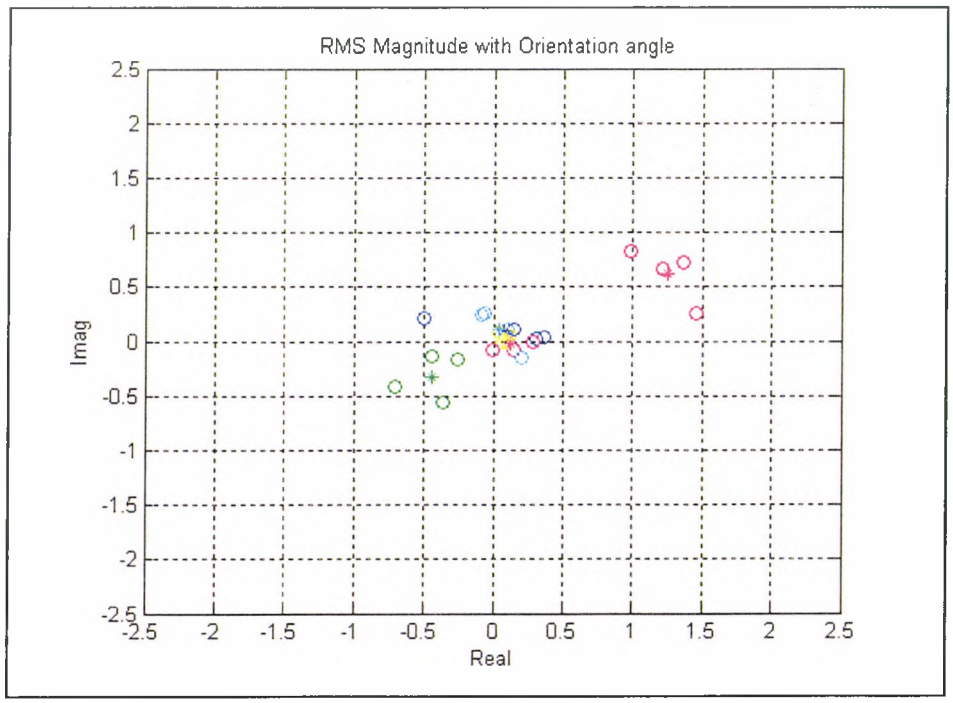


Figure A7.18 Visual grouping of gestures similar to gesture M.

Harmonic	Magnitude Average	Magnitude Std.	Orientation Angle, average	Orientation Angle, std
1	1.4234	0.1121	26.819	12.5799
2	0.5674	0.226	-145.259	16.7065
3	0.3484	0.1517	51.1561	72.8602
4	0.2296	0.0542	56.3197	67.6259
5	0.1579	0.0911	-14.9375	65.7921
6	0.0944	0.0246	15.7657	33.4328
7	0.0796	0.0263	8.6288	39.4531
8	0.0474	0.0249	-28.0281	10.4593
9	0.0293	0.0115	31.2626	50.7381
10	0.0208	0.0048	-20.4761	54.1268
11	0.0163	0.0031	-35.5368	62.6274
12	0.0165	0.0026	-3.9694	50.8193

Table A7.10 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture M. Normalised by A_p =1

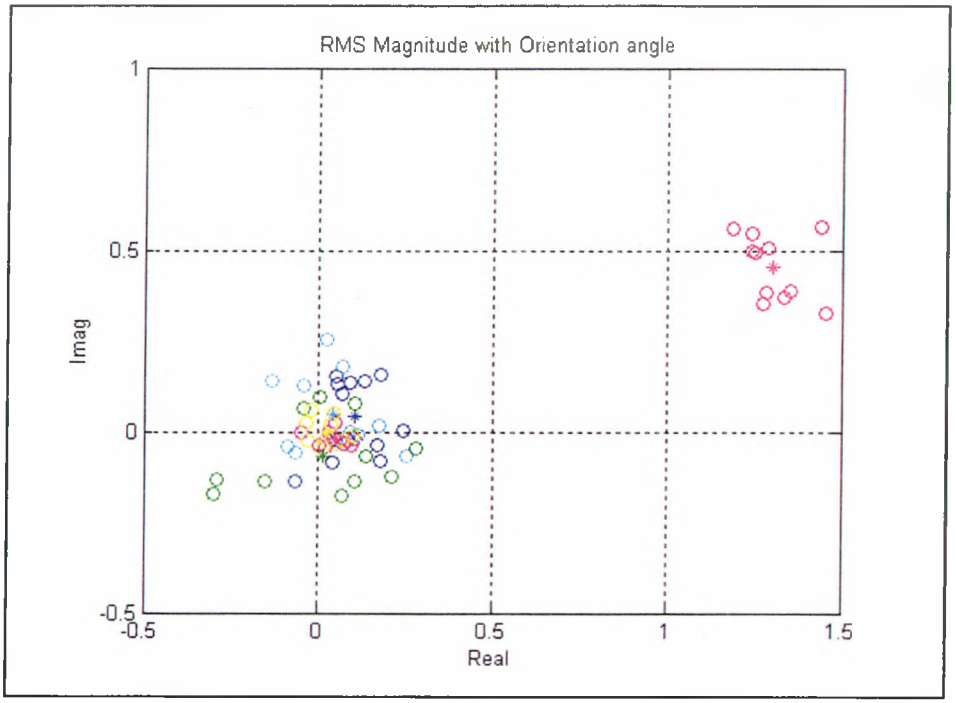


Figure A7.19 Visual grouping of gestures similar to gesture K.

Harmonic	Magnitude Average	Magnitude Std.	Orientation Angle, average	Orientation Angle, std
1	1.3859	0.0737	19.1848	4.0684
2	0.202	0.0893	-34.6788	92.5464
3	0.1715	0.0448	11.0048	60.2589
4	0.1502	0.0688	5.8319	91.9403
5	0.0688	0.0308	-36.1104	54.9963
6	0.0546	0.0203	-6.8232	83.3211
7	0.0295	0.014	24.7645	47.1823
8	0.0229	0.011	24.0979	74.062
9	0.0193	0.0103	-48.6122	81.6982
10	0.0206	0.0101	32.1101	88.6559
11	0.0186	0.0095	-37.7355	45.3002
12	0.0155	0.0084	12.6092	54.4028

Table A7.11 Magnitude and Orientation Angles (average and standard deviation) of visual grouping of gestures like gesture K. Normalised by A_p =1

4 Gesture Stimuli Sequence Experiments

A montage of images showing the positions of the first three most significant 'scm' objects (red, green and blue crosses) are shown in Figures A3.14 to A3.19.

The 'whisk' sequence gesture

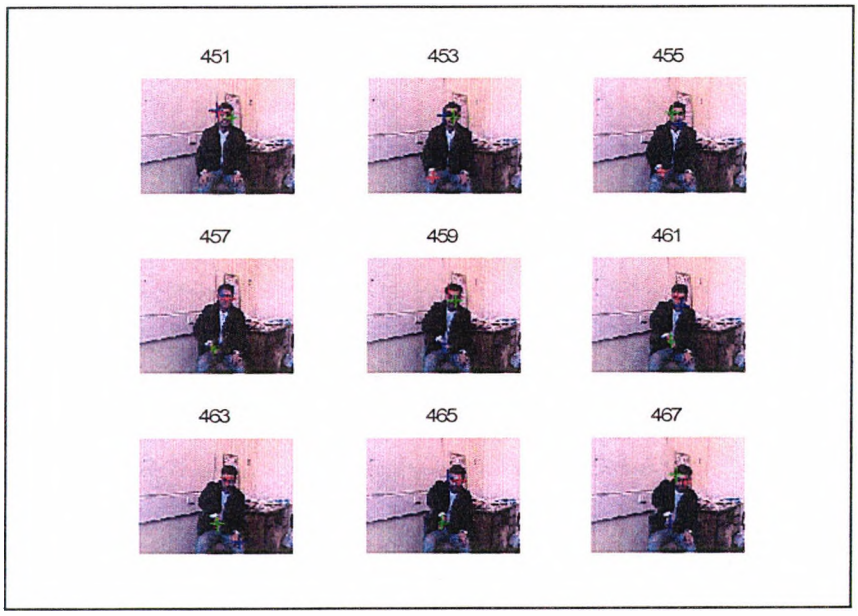


Figure A7.20 Alternate frames 451 to 457 of a 'whisk' gesturer

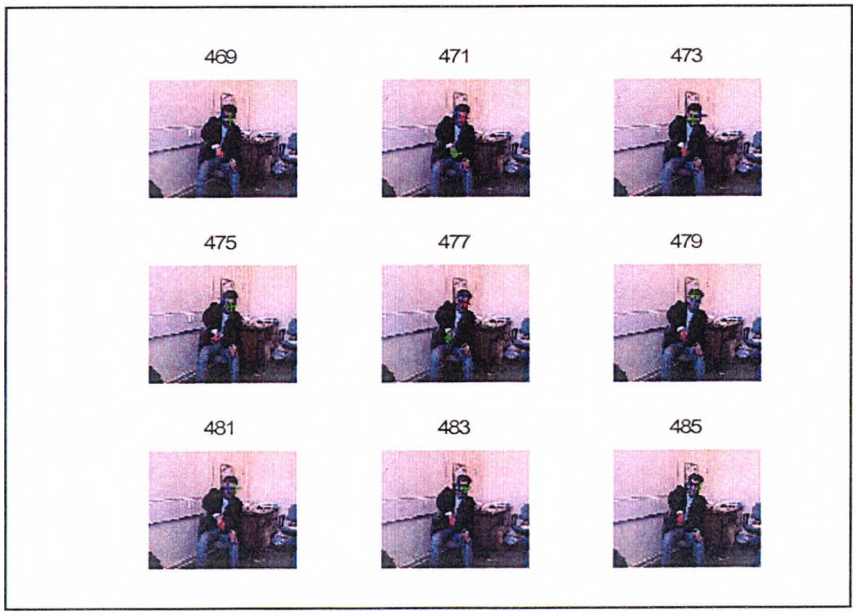


Figure A7.21 Alternate frames 469 to 485 of a 'whisk' gesturer

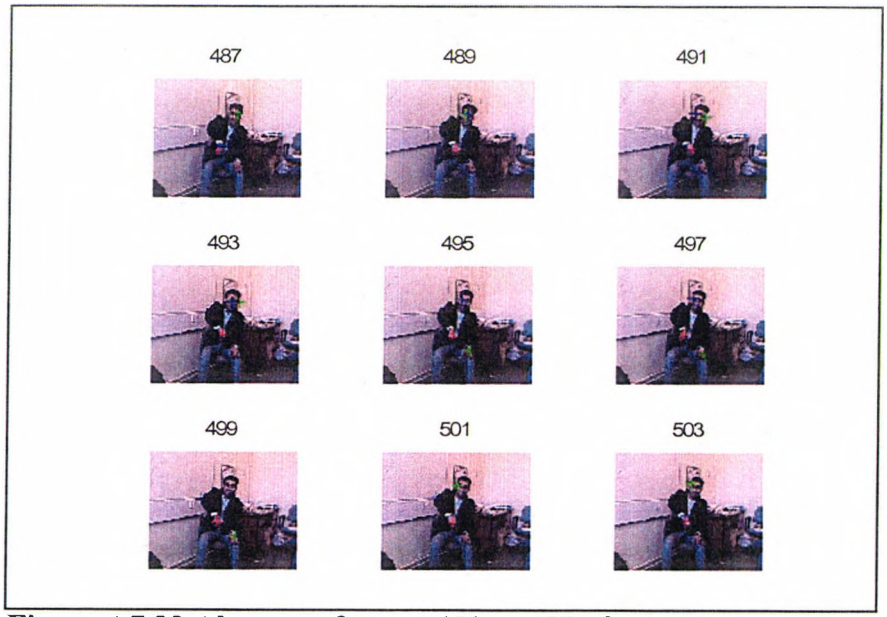


Figure A7.22 Alternate frames 451 to 457 of a 'whisk' gesturer

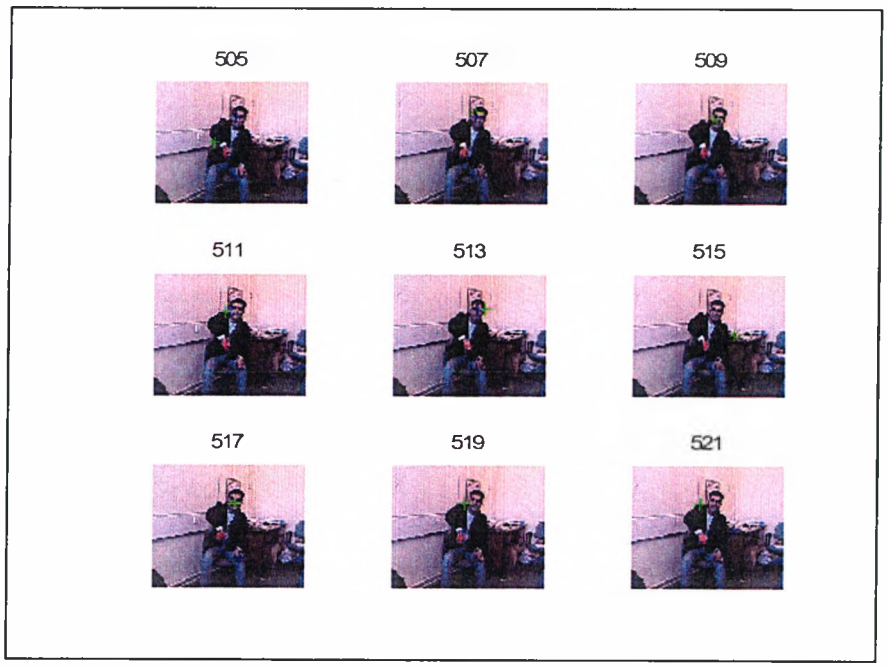


Figure A7.23 Alternate frames 451 to 457 of a 'whisk' gesturer

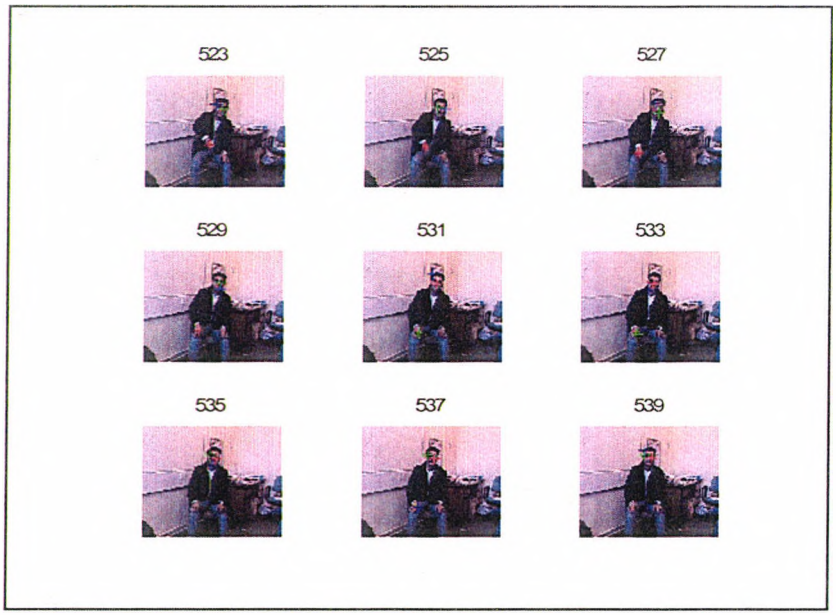


Figure A7.24 Alternate frames 451 to 457 of a 'whisk' gesturer

The 'saw-action' gesture sequence

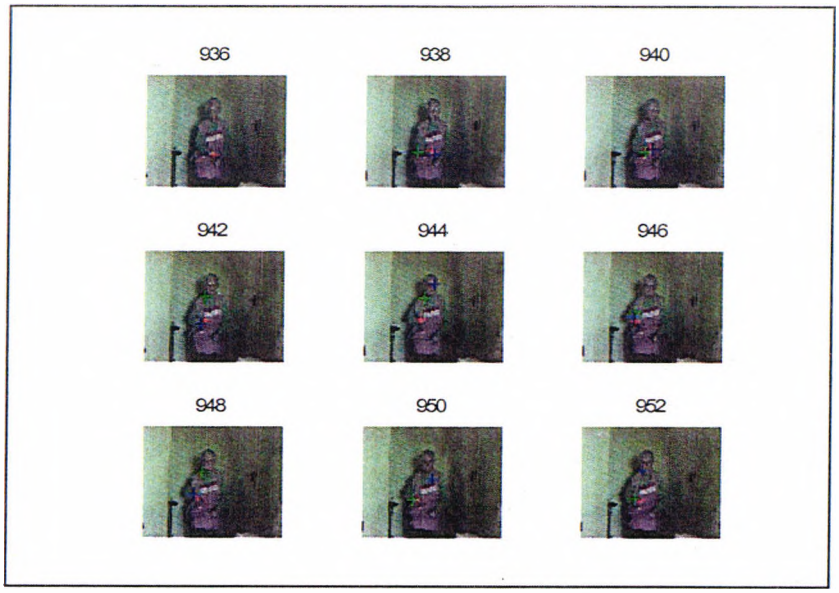


Figure A7.25 Alternate frames 936 to 952 of a 'saw-action' gesturer

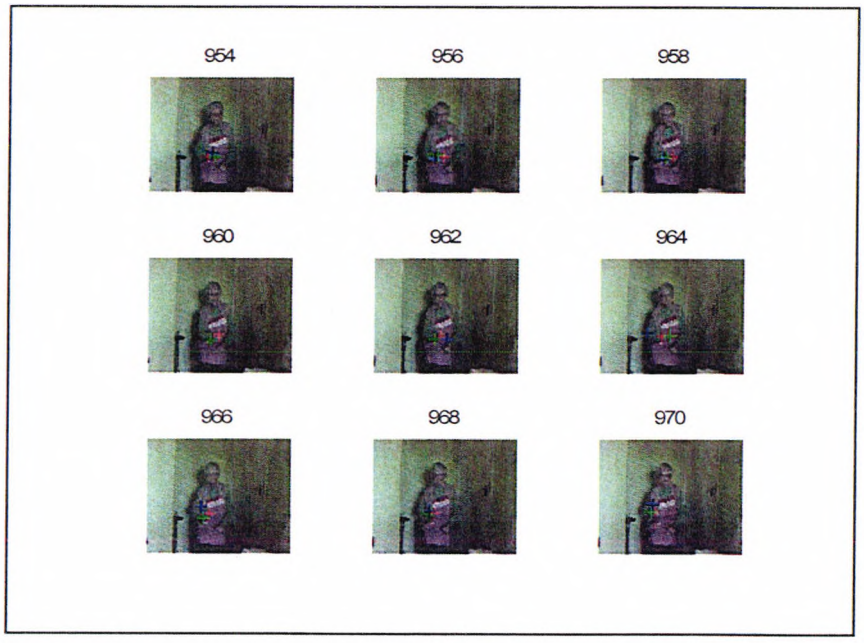


Figure A7.26 Alternate frames 954 to 970 of a 'saw-action' gesturer

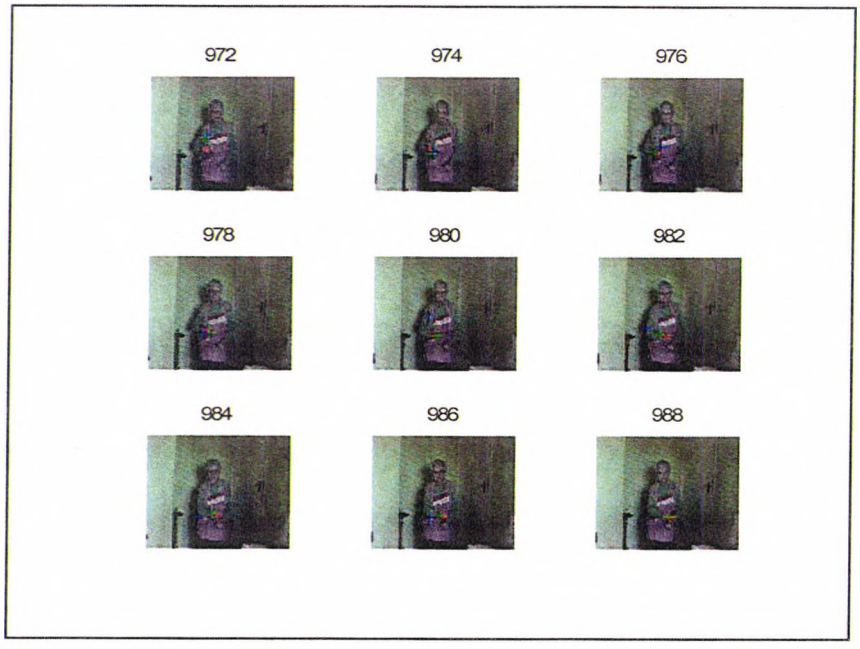


Figure A7.27 Alternate frames 972 to 988 of a 'saw-action' gesturer

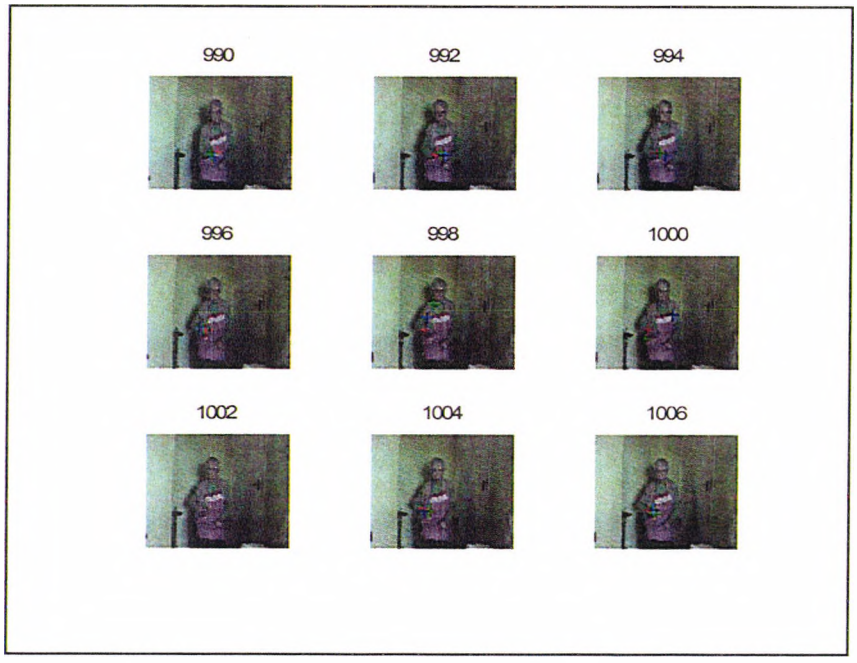


Figure A7.28 Alternate frames 990 to 1006 of a 'saw-action' gesturer

Figure A7.29 Alternate frames 1008 to 1024 of a 'saw-action' gesturer

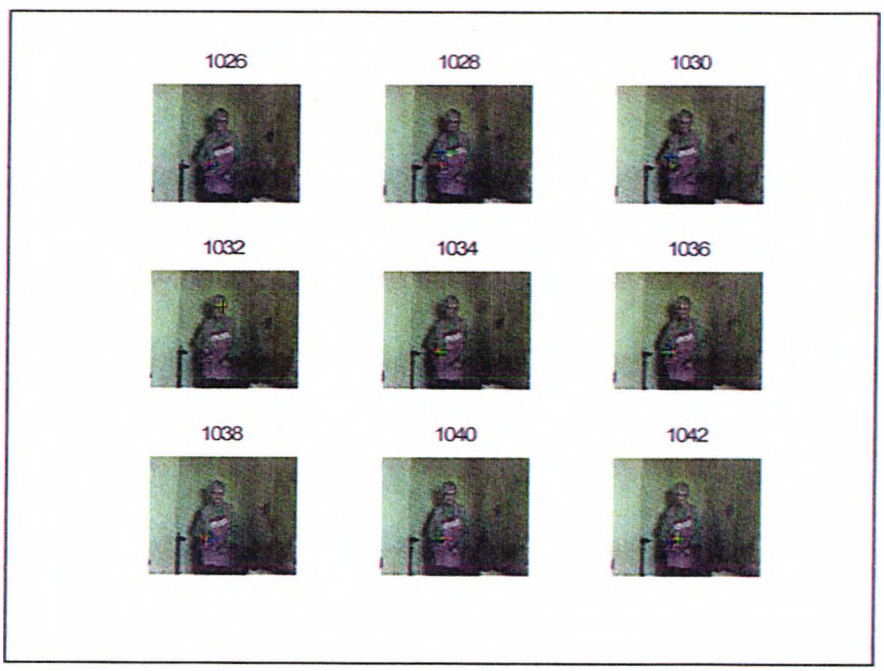


Figure A7.30 Alternate frames 1026 to 1042 of a 'saw-action' gesturer

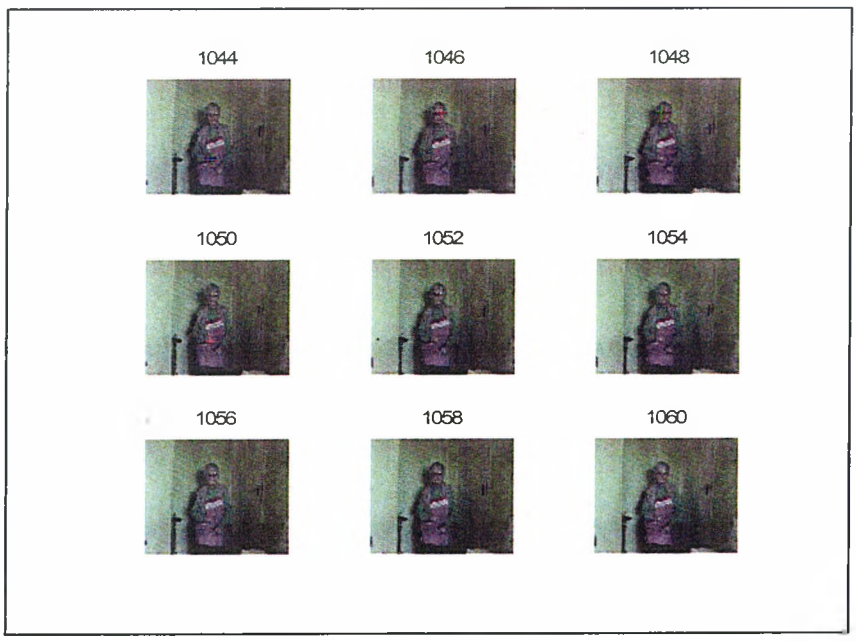


Figure A7.31 Alternate frames 1044 to 1060 of a 'saw-action' gesturer

Typical visual observations of a gesturer's response to the gesture stimuli.

The number in the column indicates the number of repetitive actions observed.

	Stimuli	Dominant hand Right	Non-Dominant Hand Left
1	Toothbrush	Right to left 1 2 times on each side	
2	Knife	4 actions	Comes up and in sympathy with right hand
3	Key	2 actions	
4	Screwdriver	10 wrist actions	
5	Hair Brush	4 actions	
6	Hammer	5 hammer actions	
7	Whisk	25 actions	
8	Saw	4 actions	
9	Bottle opener	3 actions	Left hand comes upholding virtual bottle

Table A7.12 A gesturer's response to gesture stimuli