The Active Stereo Probe: The Design and Implementation of an Active Videometrics System

Colin W. Urquhart

The Turing Institute

June 1997

Submitted for the degree of Doctor of Philosophy to:
The Department of Computing Science,
The University of Glasgow.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

Abstract

This thesis describes research leading to the design and development of the Active Stereo Probe (ASP): an active vision based videometrics system. The ASP espouses both definitions of active vision by integrating structured illumination with a steerable binocular camera platform (or head). However, the primary function of the ASP is to recover quantitative 3D surface models of a scene from stereo images captured from the system's stereo pair of CCD video cameras.

Stereo matching is performed using a development of Zhengping and Mowforth's Multiple Scale Signal Matcher (MSSM) stereo matcher. The performance of the original MSSM algorithm was dramatically improved, both in terms of speed of execution and dynamic range, by completely re-implementing it using an efficient scale space pyramid image representation. A range of quantitative performance tests for stereo matchers was developed, and these were applied to the newly developed MSSM stereo matcher to verify its suitability for use in the ASP. The performance of the stereo matcher is further improved by employing the ASP's structured illumination device to bathe the imaged scene in textured light.

Few previously reported dynamic binocular camera heads have been able to perform any type of quantitative vision task. It is argued here that this failure has arisen mainly from the rudimentary nature of the design process applied to previous heads. Therefore, in order to address this problem, a new rigorous approach, suitable for the design of both dynamic and static stereo vision systems, was devised. This approach relies extensively upon system modelling as part of the design process. In order to support this new design approach, a general mathematical model of stereo imaging systems was developed and implemented within a software simulator. This simulator was then applied to the analysis of the requirements of the ASP and the MSSM stereo matcher. A specification for the imaging and actuation components of the ASP was hence obtained which was predicted to meet its performance requirements. This led directly to the fabrication of the completed ASP sensor head. The developed approach and model has subsequently been used successfully for the design of several other quantitative stereo vision systems.

A vital requirement of any vision system that is intended to perform quantitative measurement is calibration. A novel calibration scheme was devised for the ASP by adopting advanced techniques from the field of photogrammetry and adapting them for use in the context of a dynamic computer vision system. The photogrammetric technique known as the Direct Linear Transform was used successfully in the implementation of the first, static stage of this calibration scheme.

A significant aspect of the work reported in this thesis is the importance given to integrating the components developed for the ASP, i.e. the sensor head, the stereo matching software and the calibration software, into a complete videometric system. The success of this approach is demonstrated by the high quality of 3D surface models obtained using the integrated videometric system that was developed.

Acknowledgements

I gratefully acknowledge the support for this work provided by the United Kingdom Department of Trade and Industry through its partial funding of the Active Stereo Probe project (JFIT IED3/1/2109) and its grant of a SMART award. I am also grateful to the Turing Institute, in all of its forms, for providing support and facilities for this work.

I would like to thank all of the people who have helped me during the course of this work. In particular, I would like to thank: the other Active Stereo Probe project team members, John McDonald, David Wilson, John Heng and Richard Fryer, for their assistance and advice; Joe Jin (formerly Jin Zengping) and Peter Mowforth for providing me with the original MSSM stereo matcher that started everything off; Arthur Van Hoff for his great work in developing Version III of MSSM; Professor Gordon Petrie of the Department of Geography and Topographic Science, University of Glasgow, for introducing me (sometimes painfully!) to the science of photogrammetry and for his helpful advice and thorough comments; Eddie Grant of the Department of Computer Science, University of Strathclyde, for his generous assistance at a time when I'm sure that he had much more important things to do; my parents for their unwavering support; and Gillian for her patience and understanding.

However, most of all I would like to thank Paul Siebert, not only for supervising me for far longer than he ever thought, but also for his limitless enthusiasm and constant encouragement.

Contents

Chapter 1 Introduction **1**

1.1. Recovering the third dimension 1

1.1.1. Requirement 1

1.1.2. Current methods 1

1.2. Background 5

1.2.1. Previous Work 5

1.2.2. The Active Stereo Probe 6

1.3. Objective 8

1.4. Thesis structure 9

Chapter 2 Literature review **10**

2.1. Introduction 10

2.2. Photogrammetry 11

2.2.1. Introduction 11

2.2.2. Basic principles and terminology 12

2.2.3. Early photogrammetry 15

2.2.4. Analogue photogrammetry 16

2.2.5. Analytical photogrammetry 17

2.2.6. Digital photogrammetry 19

2.2.7. Photogrammetric potential of non-metric and video cameras 19

2.3. Calibration of static video cameras 21

2.3.1. Introduction 21

2.3.2. Tsai's method 21

2.3.3. Other selected methods 23

2.4. Computer stereo vision 25

2.4.1. Overview 25

2.4.2. Passive methods 25

2.4.3. Active methods 31

2.4.4. Combined active and passive methods 34

2.5. Videometrics 35

2.6. Dynamic computer vision 36

2.6.1. The dynamic vision paradigm 36

2.6.2. Benefits of dynamic vision 37

2.6.3. Dynamic vision heads 37

2.6.4. Calibration of dynamic vision systems 39

2.7. Conclusions 42

Chapter 3 Approach 44

3.1. Objectives 44

3.2. Review of problem 46

3.2.1. Overview 46

3.2.2. Sensor system design 46

3.2.3. Stereo correspondence 47

3.2.4. Dynamic calibration 48

3.3. Approach 49

3.3.1. Overview 49

3.3.2. Performance specification 50

3.3.3. Sensor system module development 50

3.3.4. Stereo matching module development 52

3.3.5. System calibration module development 52

3.4. Performance requirements 55

3.4.1. Measurement accuracy 56

3.4.2. Working volume 56

3.4.3. Disparity limit 57

3.4.4. Measurement speed 58
3.4.5. Degrees of freedom 59
3.4.6. Speed of actuation 59
3.5. Conclusions 59

Chapter 4 Sensor system model 62

4.1. Introduction 62
4.1.1. Objective 62
4.1.2. Background 62
4.1.3. Approach 65
4.2. Kinematic model 66
4.2.1. Sensor head arrangement 66
4.2.2. Homogeneous co-ordinate transformation 67
4.2.3. The viewing transformation 68
4.2.4. The world-azimuth actuator link 70
4.2.5. The azimuth actuator-elevation actuator link 70
4.2.6. The elevation actuator-camera link 72
4.3. Imaging model 73
4.3.1. Image formation 73
4.3.2. Back projection 74
4.3.3. Stereo back projection 75
4.3.4. Epipolar line derivation 77
4.4. Software implementation 79
4.4.1. Camera intrinsic parameters module 79
4.4.2. Camera extrinsic parameters module 80
4.4.3. Simulation setup module 81
4.4.4. Simulation input module 82
4.4.5. Simulation execution module 83
4.4.6. Simulation output module 84
4.5. Conclusions 85

Chapter 5 Sensor system configuration **87**

5.1. Introduction 87

5.1.1. Objective 87

5.1.2. Stereo imaging geometry 87

5.1.3. Influencing factors 89

5.1.4. Analytical approach 91

5.1.5. Initial configuration 92

5.2. Choice of cameras 93

5.3. Choice of mean object distance 94

5.4. Determination of principal distance 96

5.4.1. Objective 96

5.4.2. Method 96

5.4.3. Working volume vs. principal distance results 99

5.4.4. Working volume vs. vergence point results 100

5.4.5. Interpretation of results 103

5.5. Determination of stereo baseline 103

5.5.1. Objective 103

5.5.2. Theory 104

5.5.3. Method 106

5.5.4. Disparity limits vs. baseline results 108

5.5.5. Disparity limits vs. vergence point results 110

5.5.6. Interpretation of results 114

5.6. Conclusions 115

5.6.1. Review of objectives 115

5.6.2. Summary of results 116

5.6.3. Conclusions 117

Chapter 6 Sensor system actuation **119**

6.1. Introduction 119

6.1.1. Background 119

6.1.2. Objective 119
6.1.3. Parameters under analysis 120
6.1.4. Absolute error and residual error 121
6.1.5. Basic approach 122
6.2. Basic analytical procedure 122
6.2.1. Step 1: generation of "actual" surface points 123
6.2.2. Step 2: introduction of parameter errors 125
6.2.3. Step 3: generation of "measured" surface points 127
6.2.4. Step 4: calculation of surface measurement error 128
6.3. Individual parameter analysis 129
6.3.1. Aim 129
6.3.2. Method 129
6.3.3. Results 130
6.3.4. Interpretation of results 134
6.4. Combined parameter analysis 134
6.4.1. Aim 134
6.4.2. Actuator configuration 135
6.4.3. Method 136
6.4.4. Results 138
6.4.5. Interpretation of results 139
6.5. Conclusions 140
6.5.1. Review of objective 140
6.5.2. Summary of results 140
6.5.3. Conclusions 141

Chapter 7 Sensor system fabrication 142

7.1. Introduction 142
7.2. Selection of actuators 142
7.3. Sensor head assembly 143
7.4. Projection system integration 144

7.5. Full system integration 148

7.6. Performance verification 149

7.7. Conclusions 150

Chapter 8 Stereo matcher development 151

8.1. Introduction 151

8.1.1. Objective 151

8.1.2. Stereo matcher requirements 152

8.1.3. Shortcomings of the original MSSM stereo matcher 152

8.1.4. Overview of MSSM development 154

8.2. Original version 155

8.2.1. Basic structure 155

8.2.2. Blurring stage 155

8.2.3. Matching stage 157

8.2.4. Windowed correlation calculation 159

8.2.5. Two dimensional sub-pixel interpolation 160

8.2.6. Implementation 160

8.3. Version II 160

8.3.1. Irregular pyramid implementation 160

8.3.2. Complete re-coding 164

8.3.3. Results 164

8.4. Version III 165

8.4.1. Overview 165

8.4.2. Implementation 165

8.4.3. Results 166

8.5. Conclusions 168

Chapter 9 Stereo matcher accuracy analysis 170

9.1. Introduction 170

9.1.1. Objective 170

9.1.2. Factors affecting stereo matcher accuracy 171

9.1.3. Approach 172

9.2. Generation of test stereograms 173

9.2.1. Disparity map synthesis 173

9.2.2. Generation of right image 173

9.2.3. Generation of left image 175

9.3. Basic performance tests 176

9.3.1. Objective 176

9.3.2. Input stereograms 176

9.3.3. Method 177

9.3.4. Results 178

9.3.5. Interpretation of results 178

9.4. Accuracy versus integer disparity 181

9.4.1. Objective 181

9.4.2. Method 181

9.4.3. Results 182

9.4.4. Interpretation of results 182

9.5. Accuracy versus sub-pixel disparity 183

9.5.1. Objective 183

9.5.2. Method 183

9.5.3. Results 183

9.5.4. Interpretation of results 184

9.6. Accuracy versus disparity gradient 184

9.6.1. Objective 184

9.6.2. Method 184

9.6.3. Results 185

9.6.4. Interpretation of results 186

9.7. Verification of disparity gradient results 186

9.7.1. Objective 186

9.7.2. Method 187
9.7.3. Results 187
9.7.4. Interpretation of results 188
9.8. Accuracy versus image noise 188
9.8.1. Objective 188
9.8.2. Method 189
9.8.3. Results 189
9.8.4. Interpretation of results 191
9.9. Conclusions 191

Chapter 10 Static calibration 193

10.1. Introduction 193
10.1.1. Background 193
10.1.2. Objective 194
10.1.3. Approach 194
10.2. Design and construction of the calibration grid 196
10.2.1. Requirements 196
10.2.2. Possible configurations 196
10.2.3. Prototype calibration grid 198
10.2.4. Replacement calibration grid 199
10.3. Target detector development 201
10.3.1. Requirements 201
10.3.2. Contour tracing 201
10.3.3. Circle finding 204
10.3.4. Matching of world and image points 207
10.3.5. Accuracy analysis 207
10.3.6. Other sources of error 211
10.4. DLT based static calibration 212
10.4.1. Introduction 212
10.4.2. Derivation 212

10.4.3. Implementation 213

10.4.4. Accuracy analysis 216

10.5. Conclusions 221

Chapter 11 System integration and final results 222

11.1. Introduction 222

11.2. Software integration and user interface design 223

11.2.1. Overview 223

11.2.2. Sensor head control 223

11.2.3. System calibration 225

11.2.4. Stereo matching 225

11.2.5. Space intersection 227

11.2.6. 3D computer graphics 227

11.3. 3D surface recovery 228

11.3.1. Objective 228

11.3.2. Method 228

11.3.3. Results 228

11.4. Conclusions 231

Chapter 12 Conclusions and further work 232

12.1. Review of thesis goals 232

12.2. Achievements 233

12.2.1. Sensor system design 233

12.2.2. Computer stereo matching 234

12.2.3. Vision system calibration 236

12.2.4. System integration 237

12.3. Subsequent and future work 237

12.3.1. Analysis of 3D surface reconstruction accuracy 237

12.3.2. Stereo matcher development 238

12.3.3. Stereo matcher analysis 239

12.3.4. Static calibration 239

12.3.5. Dynamic calibration 240

12.3.6. Dynamic vision 244

12.3.7. Further system development 245

12.4. Summary of contributions 249

References **250**

Publications **261**

Appendix A Bundle adjustment

List of figures

Figure 1.1.	The ideal stereo vision arrangement with parallel camera axes and the pin-hole camera model.	3
Figure 1.2.	The basic arrangement of a "laser striper" system.	5
Figure 2.1.	The "pin-hole" camera model.	12
Figure 3.1.	Basic components of a stereo 3D computer vision system.	45
Figure 3.2.	Preliminary specification of the working volume (dashed lines) and volume of possible vergence points (solid lines) for the ASP stereo vision system.	57
Figure 4.1.	The "parallel" stereo viewing geometry.	63
Figure 4.2.	The "symmetrically verged" stereo viewing geometry.	64
Figure 4.3.	The configuration of cameras and actuators chosen for the ASP sensor head.	67
Figure 4.4.	Co-ordinate frames for one actuated camera. Links are shown symbolically with dashed lines.	69
Figure 4.5.	Translation between the world and azimuth actuator co-ordinate frames, $\left({}^{W}X_{A}, {}^{W}Y_{A}, {}^{W}Z_{A}\right)$.	71
Figure 4.6.	Off axis wobble of the azimuth actuator co-ordinate frame, ϕ_A and ω_A.	71
Figure 4.7.	Rotation of the azimuth actuator co-ordinate frame, Φ_A.	71
Figure 4.8.	Perspective projection of a world point, **P**, onto an image point, **p**.	73
Figure 4.9.	Back projection of an image point, **p**, to a space line, **L**.	74

Figure 4.10. A stereo pair of image points, **p** and **q**, back project as space lines, **L** and **M**, that intersect at a space point **R**. 76

Figure 4.11. Formation of an epipolar line, **l**, on the image plane of Camera One corresponding to an image point, **q**, on the image plane of Camera Two. 77

Figure 4.12. Schematic of the software implementation of the ASP sensor system model showing the six main modules boxed. 79

Figure 4.13. ASP sensor system model main menu screen. 79

Figure 4.14. ASP sensor system model camera intrinsic parameter display screen. 80

Figure 4.15. ASP sensor system model camera extrinsic parameter display screen. 81

Figure 4.16. ASP sensor system model simulation setup display screen. 83

Figure 5.1. Plan view of a verged stereo vision system showing the viewing volume (shaded). 90

Figure 5.2. A Sony XC77RR-CE remote camera head unit with miniature 50mm lens. 93

Figure 5.3. A similar viewing volume achieved with a shorter object distance and a shorter principal distance (cf. figure 5.1). 95

Figure 5.4. Part of the simulation log for the working volume experiments showing the values of the intrinsic and extrinsic parameters. 97

Figure 5.5. Determination of the working volume by back projecting the image plane corner points onto the world co-ordinate planes ${}^{W}Z = 4.75\text{m}$ and ${}^{W}Z = 5.25\text{m}$. 98

Figure 5.6. The nine trial vergence points, $V_1, V_2, \cdots, V_9$, used during the experiments. 98

Figure 5.7. Graph of viewing volume width and height at the front and back planes (${}^{W}Z = 4.75\text{m}$ and ${}^{W}Z = 5.25\text{m}$ respectively) vs.

lens principal distance when the cameras are verged to (0.0, 0.0, 5.0m). 99

Figure 5.8. Viewing volume obtained for the central vergence point, V_5, $(0,0,5.0\text{m})$. 100

Figure 5.9. Viewing volume obtained for the vergence points $(0,0,4.5\text{m})$, $(0.5,0,4.5\text{m})$, $(0,0.5,4.5\text{m})$ and $(0.5,0.5,4.5\text{m})$. 101

Figure 5.10. Viewing volume obtained for the vergence points $(0,0,5.5\text{m})$, $(0.5,0,5.5\text{m})$, $(0,0.5,5.5\text{m})$ and $(0.5,0.5,5.5\text{m})$. 102

Figure 5.11. Derivation of a right image epipolar line from a left image point $\mathbf{p}_{i,0}$. See text for details of this process. 105

Figure 5.12. Plot of right image epipolars derived for the original left image point $\mathbf{p}1_0$, (377, 290), with the cameras verged to the central vergence point, (0.0, 0.0, 5.0m), and camera baselines ranging from 100mm to 1000mm. 108

Figure 5.13. Graph of absolute maximum horizontal disparity, $\delta_{x,\max}$, and maximum range of horizontal disparity, Δ_x, versus camera baseline for the central vergence point, (0.0, 0.0, 5.0m). 109

Figure 5.14. Graph of absolute maximum vertical disparity, $\delta_{y,\max}$, and maximum range of vertical disparity, Δ_y, versus camera baseline for the central vergence point, (0.0, 0.0, 5.0m). 109

Figure 5.15. Epipolar lines resulting in the opposite image from one of the image points, $pn_0, n = 1,2,\cdots,10$, for each of the nine different vergence points as indicated in the legend. 111

Figure 6.1. Two surfaces before removal of global errors. 122

Figure 6.2. Two surfaces after removal of global errors. 122

Figure 6.3. The ten "actual" virtual world co-ordinate points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$, at the extremes of the viewing volume and the virtual vergence point, $\mathbf{V}$. 125

Figure 6.4. Generation of the five "actual" world points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_5$, on the back plane of the viewing volume and their corresponding image points in Camera One, $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_5$, and Camera Two, $\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_5$. 126

Figure 6.5. The "measured" virtual world points, $\hat{\mathbf{P}}_1,\hat{\mathbf{P}}_2,\cdots,\hat{\mathbf{P}}_5$, are generated after the introduction of parameter errors by stereo back projection of the virtual image points $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_5$ and $\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_5$. 127

Figure 6.6. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. azimuth actuator rotation error, ε_{Φ_A}. 131

Figure 6.7. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. azimuth actuator rotation error, ε_{Φ_A}. 131

Figure 6.8. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. elevation actuator rotation error, ε_{Ω_E}. 131

Figure 6.9. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. elevation actuator rotation error, ε_{Φ_A}. 131

Figure 6.10. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. azimuth actuator X translation error, $\varepsilon_{w_{X_A}}$. 132

Figure 6.11. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. azimuth actuator X translation error, ε_{Φ_A}. 132

Figure 6.12. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. azimuth actuator Z translation error, $\varepsilon_{w_{Z_A}}$. 132

Figure 6.13. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. azimuth actuator Z translation error, $\varepsilon_{w_{Z_A}}$. 132

Figure 6.14. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. azimuth actuator "off-axis wobble", ω_A. 133

Figure 6.15. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. azimuth actuator "off-axis wobble", ω_A. 133

Figure 6.16. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. elevation actuator "off-axis wobble", κ_E. 133

Figure 6.17. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. elevation actuator "off-axis wobble", κ_E. 133

Figure 6.18. An actuated camera. 135

Figure 6.19. Schematic of the translations introduced into the ASP sensor head model. 136

Figure 7.1. Close-up of one of the actuated cameras. 145

Figure 7.2. The completed ASP sensor head including the actuated projection mirror. 145

Figure 7.3. The LCD projection system. 146

Figure 7.4. The 35mm slide projector used in the ASP sensor system. 147

Figure 7.5. The complete ASP sensor system. 147

Figure 7.6. Schematic of the complete ASP sensor system. 149

Figure 8.1. The basic structure of the MSSM stereo matcher. 155

Figure 8.2. Scale space images produced by the blurring stage of the original version of the MSSM stereo matcher. (The images have been scaled for display purposes.) 156

Figure 8.3. Matching is carried out between the reference window centred at (x_L, y_L) in the left band-pass image and the search window centred at (x_R, y_R) in the right band-pass image. 157

Figure 8.4. The search window is moved to a number of trial matching points in the right band-pass image. 158

Figure 8.5. The irregular pyramid structure used in Version II of the MSSM stereo matcher. 163

Figure 8.6. Irregular pyramid of images produced by the blurring stage of the Version II of the MSSM stereo matcher from the same input image as shown in figure 8.2. (The images have been scaled for display purposes.) 163

Figure 8.7. Output from Version III of the MSSM stereo matcher for a texture projected input stereogram. (The disparity and confidence images have been scaled for display purposes.) 167

Figure 8.8. Output from Version III of the MSSM stereo matcher for a naturally illuminated input stereogram. (The disparity and confidence images have been scaled for display purposes.) 167

Figure 9.1. Illustration of the stretching/shrinking effect of disparity gradients of +0.25 and –0.25 on a 7 × 7 pixel checkerboard image. 171

Figure 9.2. Example of a random noise stereogram and its disparity map. 174

Figure 9.3. The input images for the basic performance tests. 177

Figure 9.4. The output disparity maps (left) and confidence maps (right) from the basic performance tests. Note that, for display purposes, the disparity maps have been scaled from -5.0 (black) to +5.0 (white) and the confidence maps have been scaled from 0.0 (black) to 1.0 (white). 179

Figure 9.5. Graph of mean disparity error (pixels) and standard deviation of disparity error (pixels) against integer disparity (pixels). 182

Figure 9.6. Standard deviation of disparity error (pixels) plotted against sub-pixel disparity (pixels). 184

Figure 9.7. Graph of RMS disparity error (pixels) against disparity gradient (pixels / pixel). 185

Figure 9.8. The synthetic input disparity map and example output disparity maps for the signal-to-noise-ratios of 10.0dB, 5.0dB and 2.5dB. All are shown scaled from –5.0 (black) to +5.0 (white). 189

Figure 9.9. Plot of slices along row 64 of the disparity images shown in figure 9.8. 190

Figure 9.10. Graph of RMS disparity error (pixels) against image signal-to-noise-ratio (dB). 190

Figure 10.1. Two basic configurations considered for the calibration grid. 197

Figure 10.2. The prototype calibration grid. 198

Figure 10.3. The replacement calibration grid. 200

Figure 10.4. Illustration of the contour tracing process: the light and dark squares represent pixels with values that have opposite signs. 203

Figure 10.5. Illustration of the circle finding process. 205

Figure 10.6. Detection of circular targets in an image of the calibration grid. 206

Figure 10.7. Graph of RMS error in the detected centre of the circle against the standard deviation of the DOG filter used during contour detection, for a range of standard deviations of image noise. 210

Figure 10.8. Graph of RMS error in the detected centre of the circle against the contrast between the circle and background. 210

Figure 10.9. Error in the estimated world position of a circular target (of diameter 20mm) caused by perspective distortion against the relative angle of the target for various positions of the target with respect to the camera. 211

Figure 10.10. Deviation from the mean of the detected column and row positions of a target point in the right and left images. (Note that different scales were used to plot the row and column positions). 218

Figure 10.11. The mean image error vectors (plotted at 100× magnitude). 219

Figure 11.1. Schematic of the CopyCAD system. 224

Figure 11.2. CopyCAD interface to the image acquisition module. 226

Figure 11.3. CopyCAD interface to the system calibration module. 226

Figure 11.4. Input images. Top: stereo pair captured under textured illumination. Bottom: stereo pair captured under normal illumination. 229

Figure 11.5. Recovered surface model. Top: wire-frame and Gouraud shaded. Bottom: texture rendered. 230

Figure 12.1. New design of calibration object, showing target points that have been detected and identified automatically. 240

Figure 12.2. Schematic of the open-loop stage of dynamic calibration. 242

Figure 12.3. Schematic of the closed-loop stage of dynamic calibration. 243

Figure 12.4. C3D-2020™ in use at the Turing Institute. 246

Figure 12.5. Examples of 3D face models captured using a C3D-2020™ system at the 1996 British Machine Vision Conference. 247

Figure 12.6. Prototype C3D-clinical™ system undergoing trials at Canniesburn Hospital. 248

Chapter 1
Introduction

1.1. Recovering the third dimension

1.1.1. Requirement

Computer aided design (CAD) systems that allow users to design three-dimensional objects of any size or complexity using a computer have been available for many years and their use is now commonplace. A vast range of related computer systems are now also available to manipulate, visualise and analyse such three-dimensional data.

The steady growth in the diversity and functionality of computer systems that manipulate three-dimensional objects has allowed their use to become ever more widespread. They are no longer found exclusively in the design departments of engineering companies, but are beginning to find applications in areas as diverse as the entertainment industry and medicine. This proliferation has been fuelled in part by the migration of the software from mainframe computers to powerful workstations and, more recently, even to increasingly powerful personal computers.

However, the continued diversification of applications for computer systems that manipulate three-dimensional information is being stifled because the required data is readily available only for new objects derived from CAD sources. A quick and convenient method for capturing three-dimensional information from existing objects is not currently available. If it were, a plethora of new applications would be made achievable almost immediately and many existing applications could be greatly enhanced.

1.1.2. Current methods

Currently available methods for three-dimensional recovery can be divided into two broad classes: contact measurement systems and non-contact measurement systems. Contact measurement systems are all essentially similar in design, while, in contrast,

there is a wide variety of non-contact systems currently in use and under development. The basic principles behind contact measurement and two methods of non-contact measurement, close-range photogrammetry and three-dimensional computer vision, are described below.

1.1.2.1. Contact measurement

The only method of three-dimensional measurement to have gained any industrial prevalence is that of contact measurement. Even so, contact measurement systems are mainly to be found only in specialised metrology laboratories. Many varieties of contact measurement systems exist, but all employ a physical probe to touch the object under measurement. The three-dimensional position of the tip of the probe is measured accurately and hence the position of the point of contact on the object can be recorded. However, there are three distinct disadvantages of such systems: they are slow; even automated systems such as co-ordinate measurement machines must scan the object one point at a time; they are expensive; and, for many applications, contact between the probe and measured object is simply unacceptable since this may deform the surface resulting in inaccurate measurement or even cause damage to the surface.

1.1.2.2. Close range photogrammetry

Photogrammetry is the well established science of obtaining measurements from photographs. In the past, its use has been limited almost exclusively to the field of aerial survey for the generation of contour maps, and more recently digital terrain models (DTMs), from aerial photographs. However, there has been steady growth in the field of close range photogrammetry that attempts to apply photogrammetric principles to the measurement of "smaller" objects such as buildings, human bodies and industrial parts.

The basic principle behind photogrammetric measurement is fairly straight forward, given two, or more, photographs of a scene taken from different positions, it is possible, using triangulation, to determine the original three-dimensional location of any point from the location of its image in each of the photographs. For example, in the ideal case illustrated in figure 1.1, the Z co-ordinate of the point $\boldsymbol{P}$ can be determined, using equation (1.1), from the X co-ordinate of its two image points, x_1 and x_2 respectively, the principal distance of the cameras, c, and the baseline separation between the two viewpoints, B. Note that the difference in the position of the two image points due to the depth of the world point is termed *disparity* or *parallax*.

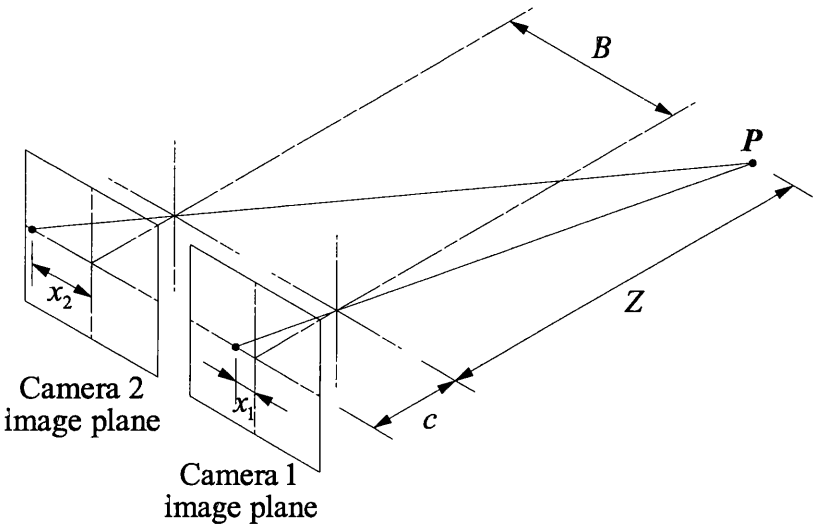


Figure 1.1. The ideal stereo vision arrangement with parallel camera axes and the pin-hole camera model.

$$Z = \frac{cB}{(x_2 - x_1)} \tag{1.1}$$

Although conventional close range photogrammetry using film cameras has proven to be extremely effective for non-contact three-dimensional measurement, it has failed to gain widespread acceptance beyond a few specialised niche applications. This can be largely attributed to the following factors: the specialised cameras usually required are very expensive; most of the techniques are extremely complex and therefore require expert knowledge; and the time taken to produce and then measure the photographs precludes many applications.

However, close range photogrammetry is potentially capable of providing an extremely dense set of surface measurements, since it is theoretically possible to derive a three-dimensional measurement for every point appearing in both (or at least two if more than two cameras are used) photographs. However most of the methods that have been developed impose the requirement that special target markers be placed on the object to be measured. These methods therefore forfeit the capability to provide a dense set of measurements since the only three-dimensional measurements that can be derived are for the sparse set of target markers.

1.1.2.3. Three-dimensional computer vision

Three-dimensional computer vision attempts to perform essentially the same task as close range photogrammetry, i.e. three-dimensional surface recovery, except using

video cameras connected directly to a computer instead of film cameras. The same basic principle of triangulation from two or more images applies. However, as well as being relatively cheap, video cameras also offer the ability to capture images into a computer both automatically and "instantaneously". Therefore, a three-dimensional computer vision system is potentially capable of fast, fully automatic (i.e. completely by machine) surface measurement.

However, there has been an enduring impediment to the fulfilment of the full potential of three-dimensional computer vision. In order to triangulate the three-dimensional position of a world point, it is first necessary to locate the corresponding image point in each of the captured images. This correspondence must then be found for a large number of imaged points if a dense set of surface measurements is required. However, it has proven to be extremely difficult to develop *passive* computer vision systems, such as those relying on stereo matchers, that are consistently successful at solving this *stereo correspondence* problem.

Many researchers have therefore chosen to develop *active* computer vision systems that attempt to ameliorate the stereo correspondence problem by projecting some form of structured light onto the scene. The simplest example of the use of structured illumination is the projection of a single spot, usually of laser light, onto the object to be measured. The task of identifying the resulting bright spot in each image is then trivial. In fact, since knowledge of the projected ray is equivalent to locating the spot in one of the images, it is possible to entirely replace one of the cameras by the illumination system. The fundamental drawback of all such systems, either using one or two cameras, is that the spot of light must be physically scanned across the surface of the object in order to build up a dense set of surface measurements.

By far the most successful active computer vision systems to date have been a class known as "laser stripers", a few of which have even been developed as commercial systems. These operate on the same principle as the "laser spot" system describe above, except that a stripe of laser light is scanned across the object rather than a single point (see figure 1.2). Each stripe thus allows a whole column of surface measurements to be derived at a time, vastly reducing the amount of scanning required. Nevertheless, scanning times are still typically of the order of a few seconds to a minute. This lengthy capture time, the intrinsic complexity of a physical scanning device, and the safety considerations involved in the use of laser light, have all limited the broad acceptance of laser stripers.

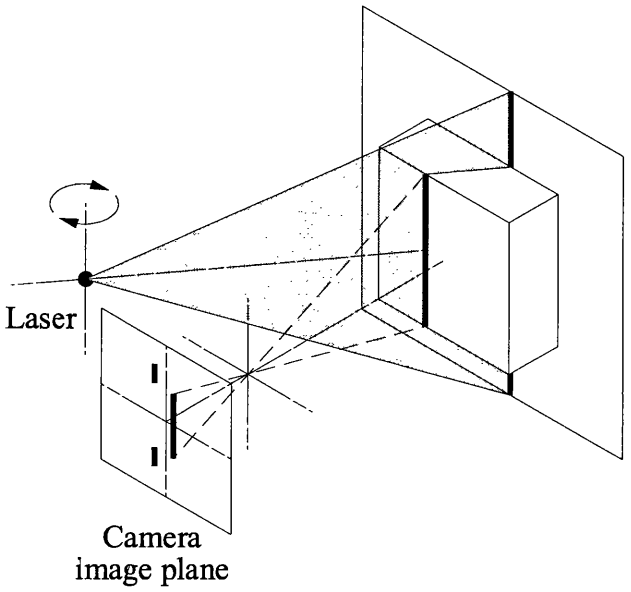


Figure 1.2. The basic arrangement of a "laser striper" system.

Considering the closeness of the fields of close range photogrammetry and three-dimensional computer vision, it is surprising that there has, until recently, been very little recognition of each of these fields by the other. Indeed, the author himself only became aware of the field of photogrammetry and its rich source of literature during the course of this work. It is apparent from both sets of literature, as described in chapter 2, that researchers working in each of these fields have often attempted to re-invent a "wheel" already well understood in the other.

1.2. Background

1.2.1. Previous Work

The principal inspiration for the work described in this thesis was a series of experiments in computer stereo vision performed by the author between April 1989 and April 1990 [Urquhart, 1989; Urquhart 1990]. For these experiments, only relatively primitive experimental equipment was available for capturing stereo images: a single CCD camera that was mounted on a piece of optical bench and translated manually between two viewpoints. A copy of the multi-scale signal matcher (MSSM) algorithm that had recently been developed by Zhengping and Mowforth [Zhengping, 1988; Zhengping, 1989] was then used to match the captured stereo images.

The resulting system was both slow and difficult to use. Painstaking care was required to position the camera correctly at both viewpoints to ensure that the resulting stereo images were scan-line aligned. Care also had to be taken to ensure that the viewed object remained still during the time required to move the camera and capture the images. Prior to stereo matching, the global disparity offset between the stereo images had to be removed manually. Even then, stereo matching itself was extremely computationally expensive: a pair of 128 × 128 pixel images would typically require at least half an hour of processing time and a pair of 256 × 256 pixel images would require several hours (using a Sun 4 workstation).

However, despite these efforts, the signal to noise ratio of the disparity maps obtained for captured stereo images was found to be disappointingly poor. This contrasted markedly with the high quality of results obtained for synthetically generated stereo pairs of "random noise" images. Investigation showed that the difference in match quality was due to image texture being much sparser, particularly at high spatial frequencies, in stereo images of natural scenes than in the synthetically generated images.

In order to overcome this problem, experiments were performed in which the viewed scene was bathed in textured light during image capture. This was successful in boosting the high spatial frequency components in the captured stereo images and resulted in dramatic improvements in the derived disparity maps [Siebert, 1990]. The technique was refined by replacing the original illumination source, an overhead projector, with a 35mm slide projector and by trialling a number of different illumination patterns. Further quality improvements were thus obtained.

1.2.2. The Active Stereo Probe

It had thus been demonstrated that the combination of the MSSM stereo matcher and textured scene illumination could produce disparity maps that were, qualitatively, at least as good as those reported using laser striping. However, unlike laser striping, one shot data capture is inherent in this combination method. It therefore offers the potential for fast and accurate three-dimensional computer vision. These factors encouraged further investigation into *texture enhanced stereo vision.*

An ambitious project was therefore proposed that would address the remaining shortcomings of the original system: the painstaking requirements of image capture, the slow speed of stereo matching and the lack of a quantitative characterisation of the system's accuracy. This proposed project gained support from the DTI and the SERC and became the Active Stereo Probe (ASP) project (IED3/1/2109). Collaboration was

arranged between researchers at BBN Systems and Technologies, The Turing Institute and The University of Strathclyde and the ASP project began in September 1990. Unfortunately BBN had to drop out of the project in April 1991.

At the time when the ASP project was proposed, a new paradigm in computer vision was emerging, that of *dynamic vision*[1]. The founding principle of dynamic vision is that a vision system should be capable of dynamically acting in accordance with information recovered from the viewed scene: it should actively look rather than just passively see. In order to support the dynamic vision paradigm, the first examples of a new generation of dynamic computer vision systems that comprised steerable stereo cameras were beginning to appear. These systems were primarily intended as test beds for forthcoming investigations into the development of sensor control reflexes such as vergence, saccadic gaze control and pursuit.

It was proposed that the ASP project should also incorporate the dynamic vision paradigm. To this end, a significant part of the project involved basic research into the development of a complete, integrated, structure for the control of sensor reflexes. Furthermore, it was intended that the vision system developed under the ASP project should be capable of accurate quantitative measurement even while operating dynamically. An example of this would be to provide the system with the ability to direct its attention to various locations of interest in the scene and derive accurate surface measurements for each. An even more ambitious example would be to track a moving object, continually deriving measurements of its position, orientation, velocity, acceleration and rotation.

Another line of investigation proposed in the ASP project was to apply the dynamic vision paradigm to active vision. The proposal was that, using one of the recently available liquid crystal display (LCD) or spatial light modulator (SLM) projection systems, it would be possible to dynamically alter the projected structured illumination pattern in accordance with what was observed in the scene. Thus, for example, a textured light pattern could be chosen to optimally enhance the viewed scene for subsequent stereo matching of the captured images.

---

[1] Most authors prefer to use the term active vision in place of dynamic vision. However, active vision is also widely used to describe the use of structured illumination. In order to avoid confusion, the term dynamic vision will continue to be used throughout this thesis to refer to the use of steerable sensors while *active vision* will refer to the use of structured illumination.

The goal of the ASP project was, therefore, extremely ambitious: to develop a stereo computer vision system that was capable of quantitative surface measurement, that incorporated the state-of-the-art in dynamic vision principles and that utilised a highly novel active illumination source and associated techniques.

1.3. Objective

This thesis describes the author's contribution to the ASP project and related research thereafter. This work was concerned with all aspects of the development of the ASP binocular sensor system and the investigation of its use for quantitative surface measurement.

System calibration[2] to recover the position, orientation and imaging characteristics of the vision sensors is essential if quantitative measurements are to be derived from stereo observation of a scene. In fixed stereo vision systems offering even modest levels of measurement accuracy, accurate system calibration is a major concern. It is argued here that the determination of sensor orientation and position becomes an issue of even greater importance if quantitative measurements of the observed scene are to be provided by dynamic stereo vision systems that employ moveable cameras.

It is proposed in this work that feedback from the camera actuators be used to facilitate calibration of the dynamic stereo vision system. The primary objective in developing the binocular sensor head for the ASP project was therefore to ensure that sensor position and orientation could be derived with sufficient accuracy using direct encoder feedback from the camera actuators. The eventual goal was then to devise and investigate a method of surface recovery that could incorporate this sensor feedback information.

Another objective was to address the shortcomings of the MSSM stereo matcher. The primary limitation, from an operational point of view, was its slow speed of execution. However, another major shortcoming was the fact that there was no quantitative information available concerning its performance characteristics. Not even essential

---

[2] In photogrammetric terminology, *camera calibration* is performed to determine the geometric characteristics of a camera and lens, while *space resection* is performed to determine the position and orientation of the cameras. In computer vision, these two operations are usually combined and termed *system calibration* or often simply *calibration.*

information such as expected matching accuracy was available. It was therefore intended to address both of these points.

The goals of this work can be broadly summarised as follows:

- to develop a dynamic sensor platform for the ASP project that would be suitable for quantitative vision tasks;
- to improve the execution rate of the MSSM stereo matching algorithm and gain an understanding of its performance characteristics by experimental analysis;
- to devise, implement and investigate a novel method for quantitative three-dimensional surface recovery appropriate for dynamic computer vision systems, such as the ASP system, that employ moveable cameras.

1.4. Thesis structure

The literature in the areas of research relevant to the work reported in this thesis is reviewed in chapter 2. Chapter 3 summarises the problems associated with performing quantitative computer vision tasks using dynamic sensors and details the approach that was adopted in this work to solve them. A computer model that was used extensively during the design of the ASP sensor head is described in chapter 4. The experimental investigations performed using this model to design the sensor head are presented in chapters 5 and 6. Chapter 7 describes the completed ASP sensor system. Chapter 8 reports the improvements made to the MSSM stereo matcher and chapter 9 describes the performance analysis experiments that were performed. The development and analysis of a method for three-dimensional surface recovery using static stereo cameras is detailed in chapter 10. Chapter 11 describes how the sensor system, stereo-matcher and three-dimensional surface recovery software were integrated to form a complete videometrics system. Finally, in chapter 12, conclusions are drawn from the obtained results and opportunities for further work are suggested.

Chapter 2
Literature review

2.1. Introduction

The goal of the work described in this thesis was to develop the Active Stereo Probe (ASP), a dynamic computer stereo vision system capable of accurate three-dimensional measurement. In order to derive accurate three-dimensional measurements from images automatically, the proposed system must address both the fundamental problems of stereo correspondence and system calibration. The stereo correspondence problem is to identify automatically the image point in each of a pair of stereo images that correspond to the same world point. The system calibration problem is to determine the parameters of the imaging system that allows the three-dimensional position of the world point to be determined from the locations of the corresponding stereo pair of image points.

Two separate communities have been researching solutions to these problems, those of computer vision and photogrammetry. However, until fairly recently, these two communities have tended to work largely independently of each other. The photogrammetric community has tended to concentrate mainly on the problems associated with system calibration while the computer vision community has tended to concentrate more on the stereo correspondence problem. Indeed, even as recently as 1992, Gruen concluded an article with the statement that "photogrammetric methodology is not used to a great extent in robot vision yet" [Gruen, 1992]. However a few computer vision researchers have succeeded in developing their own methods for system calibration. Section 2.2 of this chapter reviews relevant work from the photogrammetric field, section 2.3 describes the main methods of system calibration that have been devised by the computer vision community and section 2.4 reviews research into computer solutions to the stereo correspondence problem.

Co-operation between the fields of computer vision and photogrammetry, that was initially rather limited, has recently become increasingly close as recognition of each field by the other continues to grow. One result of this has been a steadily mounting

interest in the use of video cameras for quantitative measurement. This field of research was initially known separately as 3D computer vision by computer vision researchers and video based, or real-time, close range photogrammetry by photogrammetrists. However the unifying term of *videometrics* is rapidly gaining acceptance in both communities. The relatively small body of literature that has so far been published in this comparatively new field will be reviewed in section 2.5.

A notable distinction of the ASP system over the majority of systems in the field of videometrics is the incorporation of principles from dynamic vision. The basic theories for this new paradigm in computer vision were developed in the mid 1980s. Since then, a growing number of computer vision systems have been developed to support dynamic vision through their use of steerable cameras. Recently the problem of calibrating this new generation of dynamic vision systems has seen increasing amounts of interest. Section 2.6 of this chapter reviews the most pertinent work from this field of research.

The two main topics of research reported in this thesis, namely videometrics and dynamic vision, were relatively new at the outset of this work. However both topics have enjoyed steadily increasing amounts of attention since then and have consequently seen significant advances. Therefore the review of literature presented in sections 2.2, 2.3 and 2.4 describes predominantly work that was carried prior to that reported in this thesis, and that provided much of the original motivation. In contrast, sections 2.5 and 2.6 describe mainly more contemporary work that has been carried out since the commencement of the research detailed herein.

2.2. Photogrammetry

2.2.1. Introduction

The third edition of the *Manual of Photogrammetry* [Thompson, 1966], which is practically the official reference for the subject, defines photogrammetry as "the science or art of obtaining reliable measurements by means of photographs". (Section 2.2 of this chapter refers extensively to the fourth edition of the *Manual of Photogrammetry* [Slama, 1980] and also to personal correspondence with Professor Gordon Petrie of the Geography and Topographic Sciences Department, The University of Glasgow). Interest in the use of photographs to derive measurements began almost with the invention of the camera itself. One of the first applications of photogrammetry in the late 1800s was in the construction of maps from photographs of the ground taken from balloons. Ever since, aerial photogrammetry for mapping

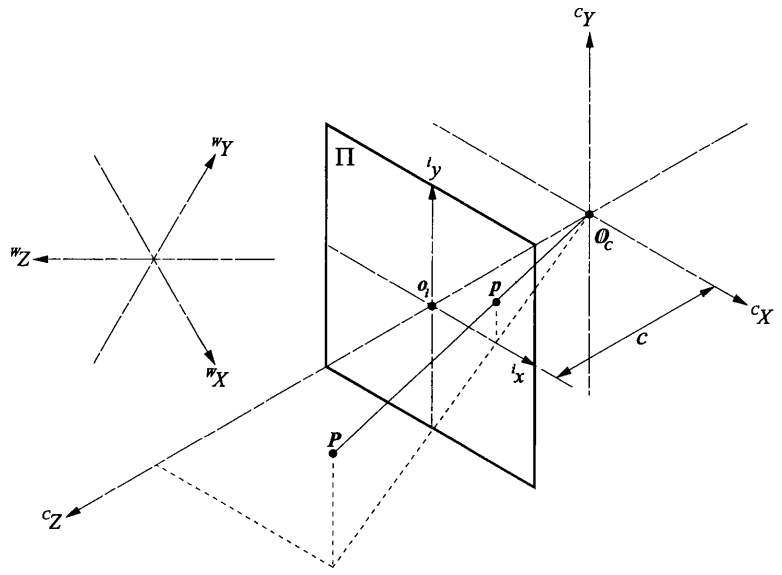


Figure 2.1. The "pin-hole" camera model.

purposes has tended to dominate the subject to such an extent that its full potential is still little recognised out with this fairly narrow area of application.

2.2.2. Basic principles and terminology

The *collinearity equations* are one of the fundamental bases of photogrammetry. They describe mathematically the relationship between the locations of three-dimensional world co-ordinate points in an observed scene and the locations of their corresponding two-dimensional image points in photographs (or more recently video or digital imagery) of the scene. Manipulation of the collinearity equations allows the three-dimensional location of points in the scene to be determined from their two-dimensional image points in two or more photographs taken from different positions.

The derivation of the collinearity equations (adapted from that given in [Slama, 1980]) begins with the definition of the "pin-hole" camera model that is shown in figure 2.1. This model comprises a *perspective centre* (or exposure centre), $\boldsymbol{O}_C$ in the figure, through which all rays of light that form the image must pass. The perspective centre forms the origin of a three-dimensional co-ordinate system for the camera whose axes are designated ${}^{C}X$, ${}^{C}Y$ and ${}^{C}Z$. This co-ordinate system is specified in such a way that the image plane, Π, is the plane parallel to the ${}^{C}X$-${}^{C}Y$ plane situated at ${}^{C}Z = c$. Where c is the *principal distance* of the camera. The point of intersection of the ${}^{C}Z$-axis and Π, $\boldsymbol{o}_i$ in figure 2.1, is termed the *principal point* of the image. The principal point

forms the origin of a two-dimensional image co-ordinate system whose axes, designated ${}^{i}x$ and ${}^{i}y$, are parallel to the ${}^{C}X$- and ${}^{C}Y$-axes respectively.

The perspective projection of a three-dimensional point $\boldsymbol{P}$, with camera co-ordinates $\left({}^{C}X_P, {}^{C}Y_P, {}^{C}Z_P\right)$, to its image point $\boldsymbol{p}$, with image co-ordinates $\left({}^{i}x_p, {}^{i}y_p\right)$is governed by the basic collinearity equations. (So called because they are based on the fact that the lines $\overline{\boldsymbol{O}_C\boldsymbol{P}}$ and $\overline{\boldsymbol{O}_C\boldsymbol{p}}$ are collinear.) It can be easily shown using the law of similar triangles that the basic collinearity equations are:

$${}^{i}x_p = c\frac{{}^{C}X_P}{{}^{C}Z_P} \tag{2.1}$$

$${}^{i}y_p = c\frac{{}^{C}Y_P}{{}^{C}Z_P} \tag{2.2}$$

In general, the image co-ordinates are not used directly. Instead image points are normally measured with respect to a co-ordinate system defined in terms of the device used to measure the photograph. In photogrammetry such a device is called a comparator. Hence a comparator co-ordinate frame can be defined with axes parallel to, but offset from, the image co-ordinate axes so that equations (2.1) and (2.2) respectively become:

$${}^{c}x_p - {}^{c}x_0 = c\frac{{}^{C}X_P}{{}^{C}Z_P} \tag{2.3}$$

$${}^{c}y_p - {}^{c}x_0 = c\frac{{}^{C}Y_P}{{}^{C}Z_P} \tag{2.4}$$

where $\left({}^{c}x_p, {}^{c}y_p\right)$ are the comparator co-ordinates of the image point $\boldsymbol{p}$ and $\left({}^{c}x_0, {}^{c}y_0\right)$ are the comparator co-ordinates of the principal point $\boldsymbol{o}_i$. Other factors, such as lens distortion can also affect the relationship between the ideal image co-ordinates and measured comparator co-ordinates of image points. Additional parameters are therefore often added to equations (2.3) and (2.4) to account for these effects. The principal distance, c, the location of the principal point, $\left({}^{c}x_0, {}^{c}y_0\right)$, and any additional lens distortion parameters are collectively known to photogrammetrists as the *interior orientation parameters* of a camera. Special *metric cameras* that are specially designed and measured so that the principal distance and location of the principal point are fixed and known accurately, are generally used in photogrammetry. The optics of these cameras are also designed to minimise lens distortion. The process of determining the interior orientation parameters is referred to as *camera calibration.*

It is not usually convenient to use the form of the collinearity equations given in equations (2.3) and (2.4) that describe world points in terms of a three-dimensional co-ordinate system fixed to the camera. It is more general, and usually more convenient, to describe world points in terms of an arbitrary world co-ordinate system. If the axes of the world co-ordinate system are ${}^{W}X$, ${}^{W}Y$ and ${}^{W}Z$, then the arbitrary transformation between world co-ordinates, $\left({}^{W}X_P, {}^{W}Y_P, {}^{W}Z_P\right)$, and camera co-ordinates, $\left({}^{C}X_P, {}^{C}Y_P, {}^{C}Z_P\right)$, for point $\boldsymbol{P}$ can be represented by the general rigid body transformation equation:

$$\begin{bmatrix} {}^{C}X_P \\ {}^{C}Y_P \\ {}^{C}Z_P \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} {}^{W}X_P \\ {}^{W}Y_P \\ {}^{W}Z_P \end{bmatrix} + \begin{bmatrix} t_X \\ t_Y \\ t_Z \end{bmatrix} \tag{2.5}$$

Where t_i are the elements of a translation vector, $\mathbf{t}$, and r_{ij} are the elements of a rotation matrix, $\mathbf{R}$, constructed by rotations about the ${}^{W}X$-, ${}^{W}Y$- and ${}^{W}Z$-axes by the Euler angles ω, ϕ, and κ respectively:

$$r_{11} = \cos\phi\cos\kappa \tag{2.6}$$
$$r_{12} = \cos\omega\sin\kappa + \sin\omega\sin\phi\cos\kappa \tag{2.7}$$
$$r_{13} = \sin\omega\sin\kappa - \cos\omega\sin\phi\cos\kappa \tag{2.8}$$
$$r_{21} = -\cos\phi\sin\kappa \tag{2.9}$$
$$r_{22} = \cos\omega\cos\kappa - \sin\omega\sin\phi\sin\kappa \tag{2.10}$$
$$r_{23} = \sin\omega\cos\kappa + \cos\omega\sin\phi\sin\kappa \tag{2.11}$$
$$r_{31} = \sin\phi \tag{2.12}$$
$$r_{32} = -\sin\omega\cos\phi \tag{2.13}$$
$$r_{33} = \cos\omega\cos\phi \tag{2.14}$$

Substituting the expressions obtained for ${}^{W}X_P$, ${}^{W}Y_P$ and ${}^{W}Z_P$ from equation (2.5) into equations (2.3) and (2.4) and dropping the prefixes *W* and *c* that designate the world and comparator co-ordinate frames respectively and the suffix *P* that identifies the point $\boldsymbol{P}$, yields:

$$x - x_0 = c\frac{r_{11}X + r_{12}Y + r_{13}Z + t_X}{r_{31}X + r_{32}Y + r_{33}Z + t_Z} \tag{2.15}$$

$$y - y_0 = c\frac{r_{21}X + r_{22}Y + r_{23}Z + t_Y}{r_{31}X + r_{32}Y + r_{33}Z + t_Z} \tag{2.16}$$

Although the representation shown in equation (2.5) is standard in computer vision and robotics, a slightly different form is more common in photogrammetric literature. In the form more commonly used in photogrammetry, the translation is carried out

before, instead of after the rotation. This gives rise to a slightly different, although equivalent, equation:

$$\begin{bmatrix} {}^{C}X_P \\ {}^{C}Y_P \\ {}^{C}Z_P \end{bmatrix} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \begin{bmatrix} {}^{W}X_P - {}^{W}X_0 \\ {}^{W}Y_P - {}^{W}Y_0 \\ {}^{W}Z_P - {}^{W}Z_0 \end{bmatrix} \tag{2.17}$$

Where $({}^{W}X_0, {}^{W}Y_0, {}^{W}Z_0)$ are the world co-ordinates of the perspective centre of the camera. The world co-ordinates of the perspective centre, $({}^{W}X_0, {}^{W}Y_0, {}^{W}Z_0)$, and the three Euler angles, ω, φ, and κ, are referred to by photogrammetrists as the *exterior orientation parameters* of a camera. The process of determining the exterior orientation parameters for a camera is referred to as *space resection.*

Substituting the expressions obtained for ${}^{W}X_P$, ${}^{W}Y_P$ and ${}^{W}Z_P$ from equation (2.17) into equations (2.3) and (2.4) and again dropping prefixes yields the more familiar form of the collinearity equations, that are often accredited to Church [Church, 1945]:

$$x - x_0 = c\frac{r_{11}(X - X_0) + r_{12}(Y - Y_0) + r_{13}(Z - Z_0)}{r_{31}(X - X_0) + r_{32}(Y - Y_0) + r_{33}(Z - Z_0)} \tag{2.18}$$

$$y - y_0 = c\frac{r_{21}(X - X_0) + r_{22}(Y - Y_0) + r_{23}(Z - Z_0)}{r_{31}(X - X_0) + r_{32}(Y - Y_0) + r_{33}(Z - Z_0)} \tag{2.19}$$

Even if the interior and exterior orientation parameters are known for a camera (following camera calibration and space resection), equations (2.18) and (2.19) do not provide enough information to determine the world co-ordinates of a point from its measured comparator co-ordinates. (This is because there are two equations in three unknowns.) However, if an additional set of equations (2.18) and (2.19) are available from a second photograph of the scene taken from a different position, then the world co-ordinates of a point can now be determined from the two corresponding sets of comparator co-ordinates. (Because there are now four equations in three unknowns.) This process, which is the basis of photogrammetry, and indeed three-dimensional computer vision, is termed *space intersection* by photogrammetrists.

2.2.3. Early photogrammetry

The origin of modern photogrammetry can be traced back to two key developments of the early 1900s. The first was presented in a series of papers by Sebastian Finserwalder that were published between 1899 and 1903 in which he established the

mathematical foundations of analytical photogrammetry[1]. The second was the invention of the first stereocomparators independently by Carl Pulfrich and Henry Fourcade in 1901. These were stereoscopic optical instruments that allowed an operator to measure the parallax (or disparity) between a pair of images. By combining these two developments it was therefore possible to perform simple photogrammetric calculations, albeit extremely slowly because only hand calculators were available at that time. The slow speed with which the required calculations could be performed prevented analytical photogrammetry from being adopted for practical applications until after the development of the digital computer (as described in sub-section 2.2.5).

2.2.4. Analogue photogrammetry

Because of the computational restrictions described in sub-section 2.2.3, most photogrammetric research between 1910 to 1955 turned to the development of opto-mechanical stereo plotting instruments. These ingenious devices, called stereoplotters, were essentially analogue computers that allowed maps to be plotted from stereopairs of photographs. They operate on the basic principle of *analogue* photogrammetry, which is to recreate a physical scale model, using a combination of mechanical and optical components, of the geometry of the cameras during the exposure of a stereopair of photographs. In analogue photogrammetry the computation of the collinearity equations required for space resection and space intersection is performed implicitly by the opto-mechanical construction of the analogue stereoplotter.

When the stereo photographs are placed in the analogue stereoplotter they are re-projected optically or mechanically so that the operator observes a virtual 3D model of the photographed scene, termed a *stereomodel.* The operator then manipulates a pair of targets, usually dots of light, one on each photograph, so that they appear as a single *floating dot* in the stereo model. If the stereoplotter is constructed and set-up with sufficient accuracy, then the actual height of the ground corresponding to the floating dot can be determined from the relative separation of the dots of light.

Although analogue methods proved to be extremely successful in aerial photogrammetry, their broader application was limited by the following factors:

---

[1] Analytical photogrammetry is "photogrammetry in which solutions are obtained by mathematical methods" [Slama, 1980].

1. Expensive metric cameras are required so that they can be simulated with sufficient accuracy in the analogue stereoplotter.

2. Only fairly simple imaging geometries can be accommodated since the operation of analogue stereoplotters is based on the parallel viewing geometry that was introduced in Chapter 1. Early stereoplotters assumed that the photographs were taken using exactly the parallel viewing geometry while later developments allowed small variations in the positions or tilt of the cameras to be "corrected".

3. Use of analogue stereoplotters is very labour intensive since all measurements must be made by a trained operator. For example the separation of the dots of light that form the floating dot is commonly fixed, effectively fixing the disparity or parallax and thus representing a fixed height. The operator can then trace out a single contour on a map by moving the floating dot in such a way that it always appears just on the surface of the stereomodel.

4. The requirements for high accuracy in the optical and mechanical components and construction of analogue stereoplotters mean that they are very expensive. These requirements also mean that they are generally very bulky and cumbersome devices.

2.2.5. Analytical photogrammetry

The development of digital computers in the early 1950s reasserted the feasibility of the analytical approach to photogrammetry, since these had the potential to perform the required calculations very quickly. The first *analytical stereoplotter* that used a digital computer to solve mathematically the relationship between photographic image co-ordinates and world co-ordinates was proposed by U.V. Helava at the National Research Council of Canada [Helava, 1957] and patented in 1964. Although interest in analytical photogrammetry was initially very limited, it began to grow slowly as the availability of mini-computers increased and their cost dropped. It also became increasingly clear that analytical stereoplotters could address many of the problems associated with the previous generation of analogue instruments. By the time of the International Society of Photogrammetry Congress in Helsinki in 1976, eight different types of analytical stereoplotters were either in operation or in development, most using a DEC PDP 11 computer.

One of the most important developments in the field of analytical photogrammetry with respect to the work described in this thesis was the development of the Direct Linear Transform (DLT) by Abdel-Aziz and Karara [Abdel-Aziz, 1971]. In the basic

form of the DLT the collinearity equations (2.18) and (2.19) are reduced to a linear relationship between world and comparator co-ordinates. The nine interior and exterior orientation parameters in equations (2.18) and (2.19) are combined into 11 dependent DLT parameters. If the image locations of at least six control points with known three-dimensional positions can be identified, then the DLT equations can be solved directly, using a computer, to determine the values of the eleven DLT parameters. This is equivalent to both space resection and camera calibration since the DLT parameters represent the interior and exterior orientation parameters. In fact, this type of camera calibration that employs known target points in the scene to be measured is referred to as *on-the-job calibration*. The DLT is described in greater detail in Chapter 10.

Karara and Abdel-Aziz later showed that the radial lens distortion parameter, k_1, could be added to the DLT to address lens distortion and hence increase accuracy [Karara, 1974]. However the linear nature of the DLT can then only be maintained if the co-ordinates of the principal point are known, otherwise an iterative solution is required to determine the eleven DLT parameters and k_1. Bopp and Krauss [Bopp, 1977] showed that the inter-dependency of the 11 original DLT parameters could lead to inaccuracy in the obtained solution. They therefore added constraints to derive an "exact solution" for the nine independent interior and exterior orientation parameters.

A different approach, which has come to be known as *bundle adjustment*, has also been developed for the solution of the collinearity equations in terms of the interior and exterior orientation parameters, e.g. [Brown, 1971; Faig, 1975; Wong, 1975; Granshaw, 1980]. The basis of this approach is to linearise the collinearity equations using a Taylor expansion and then to solve them simultaneously for a number of photographs. The main advantages of this approach are that it generates a solution in terms of the independent interior and exterior orientation parameters and that it can accommodate arbitrary numbers of additional lens distortion parameters. It is also not always necessary to place known control points in the scene at all, and for this reason such methods are often referred to as *self calibration*. However the necessity to linearise the collinearity equations requires an iterative least squares method of solution that is much more complex than the direct solution approach used in the DLT. The iteration process also requires a good initial estimate of the parameters otherwise it may be difficult or even impossible to find a solution that converges correctly.

2.2.6. Digital photogrammetry

In the early 1980s a number of universities and research institutes began to develop completely digital photogrammetric instruments. This class of instruments have become known as *digital photogrammetric workstations* or *digital stereoplotters*. These are similar to analytical stereoplotters in that the photogrammetric calculations are performed by a digital computer. However they differ significantly in that the measurement of image co-ordinates is also performed inside the computer, on digitised images (which are usually obtained by scanning the original photographs). This contrasts with analytical stereoplotters in which the image points are measured directly on the photographs themselves using traditional mono- or stereo-comparators. Perhaps the most significant advantage of the fully digital approach is that it facilitates automatic image matching, where the computer identifies stereo correspondences rather than the human operator. This will be discussed in section 2.4 of this chapter.

The first commercially available digital photogrammetry system, the Kern DSP-1, was launched at the International Society of Photogrammetry and Remote Sensing (ISPRS) Keynote Congress in 1988. Since then, several such systems have become commercially available: e.g. Intergraph IMD (Inter Map Digital), Leica/Helava DPW (Digital Photogrammetric Workstation), Zeiss Phodis, Matra T10, etc.. It has been demonstrated that these systems can provide a similar level of accuracy as analytical systems. However, most of the current range of digital photogrammetry systems require high performance Sun or Silicon Graphics workstations and are correspondingly expensive to purchase. This has slowed the widespread adoption of such systems. However a few systems, notably R-Wel's DMS system, are becoming available for less expensive personal computers. It is clear that as the price of these systems continues to fall and the photogrammetric community becomes more familiar with this new technology, digital photogrammetry will eventually replace analytical systems completely.

2.2.7. Photogrammetric potential of non-metric and video cameras

The most significant benefit of the analytical and digital approaches to photogrammetry, particularly with respect to the work described in this thesis, is the potential they offer for the use of non-metric cameras. This arises from the fact that on-the-job and self calibration methods can calculate the cameras' interior orientation parameters purely from the photographs of the object to be measured (and also possibly additional control points). Therefore, the use of metric cameras whose

interior orientation parameters have been calibrated "in the laboratory" is not necessary.

In 1976, Working Group V/2 of the International Society of Photogrammetry conducted an investigation into the photogrammetric potential of non-metric cameras. They concluded their investigation with: "the non-metric camera/computer evaluation combination has reached its fullest potential, and accuracies reaching the photogrammetric noise level have been achieved" [Faig, 1976]. The photogrammetric potential of the combination of non-metric cameras and analytical methods was further confirmed in [Kölbl, 1976] and [Nasu, 1976]. The flexibility offered by this powerful combination was found to be particularly useful for close-range applications. Metric cameras were often found to be unsuitable for close-range applications because of their expense, cumbersome size and inflexibility because of their fixed focus. Example applications include the mapping of geological structures [Brandow, 1976] and the monitoring of high speed processes [Van Wijk, 1976].

After the photogrammetric potential of non-metric film cameras was proved, several researchers within the photogrammetry community attempted to establish the potential of video cameras. The earliest attempts [Pinkney, 1978; Kratky, 1979; Burner, 1985] using vidicon-tube cameras were limited in accuracy by the instability of the imaging tubes. Solid state charge injected device (CID) or charge coupled device (CCD) sensors were found to provide much better stability and accuracies of up to 0.1 pixels were achieved [Gülch, 1984; Curry, 1985].

The low resolution of early CCD and CID sensors was quickly overcome by the rapid technological development of these solid state devices. An extremely extensive investigation into the photogrammetric potential of a modern full CCIR (Comité Consultatif International des Radiocommunications) resolution CCD camera, the Sony model number XC77CE, has shown that very high accuracies are possible [Beyer, 1992]. Investigations using other cameras, such as the Pulnix model number TM6CN [Robson, 1993], has helped to confirm to the photogrammetric community that CCD video cameras can be used for accurate three-dimensional measurement. Amin and Petrie [Amin, 1993] have also shown recently that the photogrammetric camera calibration of thermal video frame scanners is possible.

2.3. Calibration of static video cameras

2.3.1. Introduction

If the photogrammetry community has perhaps been quite slow to accept that video camera systems can be used for 3D metrology, the computer vision community, or at least some sections of it, have not. This is probably because the instability of early video cameras meant that they could not provide the level of accuracy to which photogrammetrists are accustomed, whereas the computer vision community, which, by and large, has not even been concerned about obtaining quantitative results, was happy to accept almost any level of accuracy.

The acknowledged leader of the computer vision community's attempts to develop calibration procedures for static video camera systems is Roger Tsai. (Where calibration here and henceforth in this chapter takes the meaning accepted in the computer vision community of combined on-the-job or self camera calibration and space resection.) Roger Tsai received the Best Paper Award for his groundbreaking first paper on the subject that was presented at the 1986 *IEEE International Conference on Computer Vision and Pattern Recognition* [Tsai, 1986]. This paper was later revised, extended and re-published as [Tsai, 1987]. Tsai's method is now widely used throughout the computer vision community and it is therefore described in some depth in sub-section 2.3.2 below. A selection of other methods that have been developed for the calibration of static video camera systems are described in sub-section 2.3.3.

2.3.2. Tsai's method

Tsai's method of video camera calibration is based upon the form of the collinearity equations that is given in equations (2.15) and (2.16). He stresses that the use of this form of the equations, as opposed to the form that is given in equations (2.18) and (2.19), is crucial to the development of his method. Four additional interior orientation parameters are inserted into the basic collinearity equations. These are used to model radial lens distortion and the process of digitising, inside a computer, the analogue T.V. signal that is output from a CCD video camera. The additional parameters are: k_1, the first coefficient of radial lens distortion, d_y, the vertical centre-to-centre distance between consecutive rows of CCD sensor elements, d'_x, the *effective* horizontal centre-to-centre distance between adjacent CCD sensor elements, and s_x, the horizontal scale factor. The modified collinearity equations are:

$$s_x^{-1} d_x' \left(1 + k_1 r^2\right)(x - x_0) = c \frac{r_{11} X + r_{12} Y + r_{13} Z + t_X}{r_{31} X + r_{32} Y + r_{33} Z + t_Z} \tag{2.20}$$

$$d_y \left(1 + k_1 r^2\right)(y - y_0) = c \frac{r_{21} X + r_{22} Y + r_{23} Z + t_Y}{r_{31} X + r_{32} Y + r_{33} Z + t_Z} \tag{2.21}$$

$$r^2 = \left[s_x^{-1} d_x' (x - x_0)\right]^2 + \left[d_y (y - y_0)\right]^2 \tag{2.22}$$

where (x, y) are now the column and row, or *pixel*, co-ordinates of an image point in the computer frame memory and (x_0, y_0) are the pixel co-ordinates of the principal point. Calibration requires a 3D grid of N non-coplanar target points whose world co-ordinates, $(X_i, Y_i, Z_i), i = 1, 2, \cdots, N$, are accurately known. An image of the target grid is acquired by the camera and the pixel co-ordinates, (x_i, y_i), of the image of each target point are measured. Each target point then provides a separate set of equations (2.20) and (2.21) with the world and pixel co-ordinates as knowns and the interior and exterior orientation parameters as unknowns. The goal of calibration is to determine the values of the interior and exterior orientation parameters from the system of $2N$ equations thus obtained.

In Tsai's original method, the pixel co-ordinates of the principal point, (x_0, y_0), are treated as known constants whose values are simply taken to be half the number of columns and rows respectively in the captured image. d_y and d_x', whose values can be derived from information provided by the manufacturers of the camera and frame-grabber, are also treated as known constants. A two stage method is used to determine the values of the remaining unknown interior and exterior orientation parameters. In the first stage, the Euler angles, ω, ϕ, and κ, the t_X and t_Y elements of the camera translation vector and the horizontal scale factor, s_x, are all determined by the direct solution of linear equations, i.e. without requiring any iteration.

The second stage first assumes that there is no lens distortion to derive approximate values for the principal distance, c, and the t_Z element of the camera translation vector by direct solution of linear equations. The exact values for c, t_Z and the lens distortion parameter, k_1, are then determined by solving the original set of equations (2.20) and (2.21) with all the other parameters now treated as known constants. Unfortunately the non-linear nature of equations (2.20) and (2.21) requires an iterative solution. However Tsai reports that only one or two iterations are usually required for convergence of the solution.

Tests of the accuracy of Tsai's calibration method are presented in [Tsai, 1987]. A stereo pair of cameras (Fairchild CCD 3000 cameras with 25mm Fujinon lenses) was

set up to view a very small working volume and then calibrated using the described method. The average and maximum error obtained are reported as 0.6mil and 1.8mil (approximately 0.015mm and 0.046mm) for a working volume of approximately 1.0" × 1.0" × 0.5" (25.4mm × 25.4mm × 12.7mm) and a depth range from approximately 4.0" to 4.5" (101.6mm to 114.3mm). However these results were obtained using a rather rudimentary method for detecting the pixel co-ordinates of the target points with sub-pixel precision both during calibration and during accuracy testing.

In [Tsai, 1987], Tsai also presents a slightly modified method of calibration that can be used when the horizontal scale factor, s_x, is known *a priori* and can hence be treated as a known constant. The advantage of this method, besides being slightly simpler, is that a single plane of target points can be used as long as it is not parallel to the image plane of the camera. In a later paper [Lenz, 1988], Lenz and Tsai describe a method for determining s_x. This requires that the low-pass anti-aliasing filter be removed from the camera and subsequently that the horizontal spatial frequency spectrum of a captured image be analysed. A number of methods for accurately determining the pixel co-ordinates of the principal point, (x_0, y_0), are also presented in the paper. Alternative methods for determining s_x are given in [Penna, 1991; Bani-Hashemi, 1991].

2.3.3. Other selected methods

In 1989, Tsai [Tsai, 1989] published an extensive review of all the methods of video camera calibration that had been reported at that time. However for almost all applications, Tsai recommends the use of his own algorithm. Since this review paper was published, a new class of methods has been developed. These use the estimated position of vanishing points to determine the camera's interior and exterior orientation parameters, e.g. [Caprile, 1990; Echigo, 1990; Wang, 1991]. However these methods have all proven to be much less accurate than Tsai's method and do not seem to offer any advantage over existing methods.

Weng, Cohen and Herniou [Weng, 1990; Weng 1992] present an example of a class of methods that can be used to compensate for large amounts of lens distortion. The presented method includes four parameters for modelling decentering and thin prism distortion in addition to the radial lens distortion parameter included in Tsai's method. They use a three step iterative procedure that decouples the refinement of estimates for the lens distortion parameters from the other non-distortion interior and exterior orientation parameters. The first step uses a close form linear solution of the collinearity equations, apparently similar to the DLT, that neglects all lens distortion

terms to derive an initial estimate for the non-distortion parameters. Step two refines the current estimates of the non-distortion parameters with the distortion parameters fixed. Step three refines the current estimates of the distortion parameters with the non-distortion parameters fixed. Steps two and three are repeated for a certain number of iterations.

Weng, Cohen and Herniou report that the inclusion of the radial distortion parameter, k_1, (as in Tsai's method) significantly increases the accuracy of the calibration. They report an increase in accuracy of approximately a factor of three over that obtained with no lens distortion parameters for stereo cameras fitted with 8mm lenses. The addition of their four other distortion parameters improves the accuracy by a further factor of approximately 1.25. However it appears that the accuracy improvement obtained by the inclusion of *all five* distortion parameters for the calibration of the same stereo cameras fitted with 25mm lenses is probably less than a factor of approximately 1.4. It therefore seems worthwhile implementing this complex method only if large amounts of lens distortion are suspected.

An interesting method that has been published more recently is Fan and Yuan's DLTEA-II algorithm [Fan, 1993]. This paper is particularly interesting because the DLTEA-II algorithm is based upon the basic DLT [Abdel-Aziz, 1971] that is used in the work described in Chapter 10 of this thesis. As usual, the method uses as input the measured pixel locations and known 3D world co-ordinates of a number of target points from a 3D target grid. The basic distortion free DLT is applied initially to determine the eleven DLT parameters. An iterative procedure is then applied to determine the values of one radial and two tangential lens distortion terms and to refine the estimates of the DLT parameters. As a final step, a third-order polynomial is fitted to the image residuals of the target points that remain after the iterative calibration procedure.

In their paper, Fan and Yuan compare the accuracy achieved using their DLTEA-II with other variations of the DLT. The experimental set-up consisted of: a pair of CCD cameras fitted with 25mm lenses, a baseline of approximately 420mm and an object distance of approximately 0.88m. The DLTEA-II algorithm resulted in RMSE values of only ±0.0198mm, ±0.0345mm and ±0.0794mm in X, Y, and Z respectively. The basic DLT without lens distortion compensation provided much less accuracy: ±0.1098mm, ±0.0938 and ±0.1826mm respectively. However the addition of the k_1 radial lens distortion parameter as described in sub-section 2.2.5 increased the accuracy to ±0.0435mm, ±0.0369mm and ±0.0911mm respectively.

2.4. Computer stereo vision

2.4.1. Overview

The quest to develop robust, fully automatic, computer algorithms to solve the stereo correspondence problem has been one of the major goals of the computer vision community for the last twenty years. Two main approaches to the problem have emerged: passive methods that attempt to find correspondences in a stereo pair of images of a scene solely from the information naturally contained within the images; and active methods that seek to avoid the problems encountered with the purely passive approach, by projecting some form of structured light onto the scene. However some methods have also been explored that combine the use of special illumination to enhance the performance of a passive stereo correspondence algorithm. The basic principles of all three of these methods and examples of systems that employ them are described below.

2.4.2. Passive methods

2.4.2.1. The MPG stereo algorithm

Almost all successful research into passive methods of solving the stereo correspondence problem has been influenced to some extent by the pioneering research of the late David Marr [Marr, 1982]. Based upon their earlier investigations of the human visual system, Marr and Poggio published a "computational theory of human stereo vision" [Marr, 1979]. This was later implemented as a computer algorithm by Grimson [Grimson, 1985] and has become known as the MPG (Marr, Poggio and Grimson) stereo algorithm.

The MPG stereo algorithm employs a coarse-to-fine matching strategy designed to reduce the possibility of false matches while still accommodating a large range of disparities. The input stereo pair of images are convolved with a series of four octave separated $\nabla^2 G$ (Laplacian of Gaussian) functions. Zero-crossings are detected in the coarsest stereo pair of images, i.e. the stereo pair resulting from convolution with the largest $\nabla^2 G$ function. Every point on a zero-crossing in the left image is matched with a point on zero-crossings in the right image based upon the sign change and local orientation of the zero-crossing. The maximum disparity range is chosen in accordance with the size of the $\nabla^2 G$ function to ensure that each zero-crossing has a very low probability of a false match with the same sign change.

The disparities obtained by matching the zero-crossings from the coarsest filtered stereo-pair become initial search offsets when the zero-crossings from the second filtered stereo pair are matched. Convolution with this smaller $\nabla^2 G$ function results in more zero-crossings. However this is balanced by an associated reduction in the maximum disparity range and hence a low probability of false matches is maintained. This disparity refinement process is repeated by matching the two remaining filtered stereo pairs of images in turn. The number of zero-crossings, and hence matches, increases while the maximum disparity range decreases with each smaller $\nabla^2 G$ function. Thus the large maximum disparity range obtained from the coarsest scale is combined with the greater detail obtained from the finest scale. This combination would not be possible without the use of a coarse-to-fine, or *scale space*, matching strategy.

2.4.2.2. Feature based stereo algorithms

The MPG stereo algorithm is widely recognised as being one of the first in a class of passive stereo matching approaches known as *feature based.* These are referred to as feature based because only a sparse set of image features, such as zero-crossings in the MPG stereo algorithm, are detected and matched. The advantage of the feature based approach is that there is only a small number of left image features to be matched, and there are only a few candidate right image matches for each. This sparse set of features can therefore be matched relatively quickly and with little ambiguity. Some notable examples of the feature based approach to passive stereo computation are described below.

Bernard and Thomson [Bernard, 1980] developed a single scale method that uses an "interest operator" to select a sparse set of distinguishable image points in the input stereo pair of images. Each interest point in the left image is termed a "node". All interest points in the right image that are within a given distance of the position of a node are considered candidate matches, or "labels", for that node. A match probability is calculated for each of these candidate matches on the basis of the sum of squared difference (SSD) of 5×5 pixel regions in the left and right images centred on the node and label points respectively. An iterative algorithm is applied to the match probabilities for all candidate matches between nodes and labels to choose the final match for each node. A similar feature based approach is described in [Prager, 1983].

Other techniques, such as those described in [Ohta, 1985; Medioni, 1985; Ayache, 1985] have been developed that are more similar to the MPG algorithm in that the features to be matched are edges, not single interest points. One of the most elaborate derivatives of the MPG algorithm has been developed by Hoff and Ahuja [Hoff,

1989]. Zero-crossings are detected and matched in a scale-space manner similar to that of the MPG algorithm except that the two coarsest scale images are subsampled by factors of 4 and 2 to form an image pyramid. Matches are established from the left image to the right and from the right to left and those that are found to be inconsistent are rejected. A complex surface interpolation scheme is included in the method to produce a piece-wise smooth surface map from the sparse set of consistent matches found at each scale. For a more complete review of basic feature based techniques, see [Dhond, 1989].

Although the original MPG approach was based firmly on neurobiological evidence and signal processing theory, most other feature based approaches are not so well founded. They tend only to perform robustly in a small range of application domains that result in imagery that is rich in the type of features that are to be matched. However, the main drawback of all feature based approaches to the stereo correspondence problem is the inherently sparse disparity data that they produce. This is particularly unacceptable in applications such as human body metrology where the surface to be measured is both smooth and complex. It therefore cannot be accurately reconstructed by interpolation from a small number of sparse points.

2.4.2.3. Area based stereo algorithms

A second class of stereo matchers, termed *area based*, have been developed that provide a disparity estimate for every pixel in the left image of a stereo pair. This approach therefore does not suffer from the problem of sparse data that is inherent in feature based approaches. However the vast number of potential matches that result makes it very difficult to identify the "correct" match for each pixel and avoid false matches that are locally optimal but globally incorrect. The difficulty of this problem has meant that only a few passive area based algorithms have been demonstrated successfully.

The few area based algorithms that have been successful almost exclusively use some form of scale-space, coarse-to-fine, or spatial frequency decomposition, matching strategy similar to the one first used in the MPG algorithm. Such strategies are usually based upon the convolution of the input images with $\nabla^2 G$, or similar, functions that reduce in size, and that therefore increase in the spatial frequency of their pass band, at each scale. Hence the initial scales match only the low spatial frequency components of the input images to derive estimates of disparity that are globally "correct" but locally imprecise. Subsequent scales steadily refine the local disparity estimate by matching higher and higher spatial frequency components of the input images.

Anandan's PhD thesis [Anandan 1987] presents an excellent description of one of the best purely area based algorithms. Although this algorithm is aimed at the more general problem of recovering shape from motion or optical flow-field calculation, it is still applicable to the more particular case of stereo correspondence[2]. Burt's Laplacian Pyramid [Burt, 1983] is applied to each input image. This results in a set of images whose spatial frequency content increases in octave steps and whose size doubles from scale to scale, or level to level, in the pyramid.

At each scale, starting with the smallest images in the pyramid, a Gaussian weighting window is successively centred at every pixel in the left image. For each position of the window in the left image, the SSD is calculated between it and a similar window that is centred at each of nine candidate match points in the right image. The candidate match point that results in the lowest SSD is chosen as the output match for that particular left image pixel at the current scale. The grid of 3×3 neighbouring pixels centred at the output match point from the previous scale are used as the nine candidate match points for the current scale. Note that the image reductions used in the construction of the Laplacian Pyramid mean that the separation of the candidate match points at a particular scale is effectively N pixels, where N is the size reduction at the current scale. The separation of candidate match points therefore halves at each scale, until at the final scale, where full resolution images are matched, the separation is a single pixel. The result is a robust stereo matcher that can match to a resolution of a single pixel over a maximum disparity range of several pixels.

One problem associated with area based stereo matchers is the huge computational burden that they impose. Anandan does not explicitly state the processing time required by his algorithm. However the fact that he describes in detail a version suitable for implementation on a massively parallel SIMD (single instruction, multiple data) computer, alludes to the fact that it runs slowly on conventional computers. This is supported by Bernard and Fischler who have implemented a similar, although probably less efficient, algorithm that produces a dense disparity map from a 1024×1024 pixel stereo pair of images in approximately eight minutes using a 4096 processor Connection Machine [Bernard, 1990]!

---

[2] Optical flow results in image displacements in two dimensions, while stereo matching is often considered to result in image displacements in only one dimension. However, unless the epipolar geometry of a stereo pair of images is known exactly, which is unlikely, matching of two dimensional image displacements is also required to determine stereo correspondence accurately.

Another deficiency of Anandan's image matcher that is common to most area based or feature based stereo matchers is that they can only derive integer pixel estimates of image displacement. However, the disparity encountered in stereograms of natural scenes is not actually quantised in this way but is a continuously varying function over the images. (Strictly speaking it is a piece-wise continuous function with possible jumps or breaks in disparity due to depth discontinuities, e.g. between a foreground object and the background.) Therefore the accuracy of 3D surface measurements derived from the disparity estimates provided by such algorithms will always be limited by their integer pixel resolution.

One area based stereo-matcher whose resolution is, almost uniquely, not limited by integer pixel quantisation is the Multiple Scale Signal Matcher (MSSM) stereo matcher. This was developed by Zhengping (Jin) and Mowforth at the Turing Institute [Zhengping, 1988; Zhengping, 1989]. The MSSM algorithm is similar to Anandan's; although, like the MPG algorithm, its original version does not employ image pyramids in its scale space image decomposition scheme.

In the MSSM algorithm, matching at each scale is performed by situating the right image window at a large number of *sub-pixel* candidate match points. This is achieved by calculating the values of the window elements using bilinear interpolation. The high matching sensitivity required by the small sub-pixel steps used in the search strategy is provided by calculating the full variance weighted correlation at each candidate point. (As opposed to, for example, the more approximate but less computationally expensive SSD that is calculated in Anandan's algorithm.) The greater complexity of the MSSM algorithm means that it is even more computationally expensive than comparable integer disparity area based algorithms. The original MSSM stereo matcher and the further developments that have been made to improve its performance are described in Chapter 8.

2.4.2.4. Combined feature and area based stereo algorithms

Some researchers have developed passive stereo matching algorithms that combine elements of area based and feature based techniques. Cochran and Medioni's [Cochran, 1989] approach is to first apply a modified local variance operator to the input stereo images and then match them using a single scale area based stereo matcher. The area based matcher calculates disparity from the peak correlation score between windowed regions of the left and right images. The left image window is held stationary at each pixel while the right image is scanned horizontally in single pixel steps. A complex post processing operation is applied to the disparity map output by the area based matcher to enforce a smoothness and ordering constraint. Edges are

then extracted from the left intensity image using the Canny Edge Detector [Canny, 1986] and used as cues to possible depth discontinuities. Finally an adaptive smoothing process that preserves the detected depth discontinuities is applied to the disparity map.

Day and Muller [Day, 1989] present a different approach that first uses a feature based technique followed by an area based one. The goal of the approach is to use Otto and Chau's single scale area based algorithm [Otto, 1989] since it can potentially derive a dense map of sub-pixel disparity estimates. However Otto and Chau's algorithm has a pull-in range that is limited to approximately two pixels. Hence it is not, by itself, capable of matching the much larger disparities that are present in most stereo images. The adopted solution is to first use Bernard and Thomson's integer disparity feature based stereo matcher [Bernard, 1980] to derive a coarse estimate of the disparity for a small number points. These points are then used as seed points for Otto and Chau's algorithm that can first refine the disparity for the seed points and then sheet grow a dense disparity map out from them.

2.4.2.5. Methods from the field of photogrammetry

Soon after the initial development of analytical stereoplotters, photogrammetrists became interested in automating the photogrammetric process. Therefore, the problem of determining automatically the disparity (or parallax) between stereo pairs of images became an important area for research and development in the photogrammetric community. Early solutions consisted of optical, electronic or opto-electronic correlators coupled to the viewing optics of the stereoplotter by means of beam splitters [Slama, 1980, pp. 718-721]. Later, as the technology became available, CCD video cameras were used to image the stereo photographs. These images were then digitised and correlated using a digital computer, e.g. [Ackermann, 1984; Bernard, 1986]. With the more recent development of digital photogrammetric workstations, in which the images are already stored in digital form, automatic image matching by computer has become more common and is now a standard feature on most commercially available systems.

Although research into stereo matching in the fields of photogrammetry and computer vision has developed from different origins, the current state-of-the-art in both fields appears to be area-based matching algorithms. It seems that the methods developed by photogrammetrists are less sophisticated than those developed by the computer vision community in terms of the image processing techniques that are employed. However, unlike in computer vision, where stereo matching is generally seen as an independent process, photogrammetrists have tended to integrate the stereo matchers that they have

developed into a complete photogrammetry system. The knowledge of the imaging geometry (e.g. the position of the epipolar lines) that this makes available to the stereo matcher can greatly constrain the problem, improving the overall results obtained.

2.4.3. Active methods

2.4.3.1. Introduction

The above discussion indicates that highly sophisticated and computationally intensive methods are required if stereo correspondences are to be found with even moderate accuracies using a purely passive approach. An alternative approach of active vision has therefore been developed that seeks to ameliorate, or even avoid completely, the problems encountered with the purely passive approach by projecting some form of structured light onto the scene. As a rule, active methods rely inherently upon the geometry of the projected light pattern to establish stereo correspondences. In fact, only one camera is often used, in which case correspondences are established between points in an image from the single camera and the projected pattern itself.

A diverse range of different active techniques for establishing stereo correspondences has been developed. There are too many for a complete review in this thesis, particularly since active methods are not directly within the scope of this work. However four typical approaches: *laser striping*, *Moiré fringing*, *temporal modulation* and *stereo scene coding* will be described briefly in this section. Particular reference will be made to their limitations by comparison with passive techniques. Besl [Besl, 1988] and Jarvis [Jarvis, 1983] both present more extensive reviews of active methods, as does McDonald in his PhD thesis [McDonald, 1994].

2.4.3.2. Laser striping

Perhaps the simplest and most commonly used active technique is laser striping, as described in Chapter 1 and also e.g. by Johannesson [Johannesson, 1993]. The principal drawbacks of this approach are its intrinsic complexity and lengthy capture time. These arise from the requirement to either scan precisely the laser stripe across the surface to be measured or move the surface precisely through a stationary laser stripe. In addition, the problems associated with accurately calibrating the laser source, i.e. determining the equivalent of its interior and exterior orientation parameters, have made it difficult to obtain accurate quantitative 3D measurements. However Trucco [Trucco, 1994] has recently presented a direct calibration method for laser striping systems that can provide a mean depth accuracy of between ±0.100mm and ±0.167mm (depending on the material of the surface being measured) over a

depth range of 150mm. The safety considerations arising from the use of laser light have also limited the broad acceptance of laser stripers.

2.4.3.3. Moiré fringing

Moiré fringing is an active vision technique that can potentially provide surface measurements without the requirement to solve either the stereo correspondence problem or the full system calibration problem. Light is projected through a special grating onto the surface to be measured. When the surface is then imaged through a second identical grating, a pattern of contours, called Moiré fringes, appears in the captured image. These contours correspond to fixed iso-height contours on the imaged surface, with a constant change in height between adjacent contours. The Moiré fringing technique therefore offers the potential advantage of providing direct 3D measurement from a single image.

However the depth resolution of Moiré fringing is severely limited when only a single set of fringes is used. In order to obtain sufficient resolution for most applications, one of the gratings must be moved in small increments, thus changing slightly the height of the Moiré fringes. Much higher height resolution can then be obtained by combining the contours from a sequence of images in which one image is captured for each position of the grating. However this clearly removes the advantage of single image capture and thus reduces the applicability of the technique to the measurement of only inanimate objects. It also requires the added complexity of a precision mechanism to move the grating.

There are other disadvantages of the Moiré fringing technique. It cannot determine the direction of the slope of a surface, i.e. whether it is concave or convex, without additional information from a different source. It cannot resolve even moderately steep slopes because the fringes become too closely packed to be accurately distinguished from each other. In common with most techniques based on pattern projection, the presence of local image texture on the surface to be measured makes it difficult or even impossible to determine the exact position of the pattern. A description and comparison of the performance of a Moiré fringing system and a laser striping system developed at NEL (the National Engineering Laboratory) is given by Marshall [Marshall, 1991].

2.4.3.4. Temporal modulation

Temporal modulation, e.g. [Sato, 1987], is another active technique that has attracted considerable attention for solving automatically the correspondence problem. The

basic principle of this technique is to project a sequence of black and white vertical stripes of different widths onto the surface to be measured. The sequence is designed so that the combination of projected black and white stripes uniquely identifies the horizontal position of each column within the projected pattern. Therefore the correct reconstruction of each frame of the projected sequence in images of the surface establishes directly the horizontal correspondence between imaged points and the projected pattern. It is theoretically possible to use this projector-to-camera disparity map in the same way as a left-camera-to-right-camera disparity recovered using passive stereo correspondence techniques to determine the 3D shape of the imaged surface.

However, difficulties are encountered with the calibration of the projection system that are analogous to those previously described for the calibration of laser striping systems. Problems that are similar to those experienced in Moiré fringing are also encountered in the presence of strong intrinsic texture on the surface to be measured and with limited depth resolution. The requirement to project a sequence of accurately aligned patterns means that temporal modulation can only be used effectively to measure inanimate objects. The requirement that each pattern be accurately aligned inside the projector makes it difficult to implement the technique using a conventional slide projector. However it is relatively simple to achieve the required stability using more recently developed liquid crystal projectors although these currently exhibit low illuminating power and thus exacerbate the problem of detecting robustly the projected pattern.

2.4.3.5. Stereo scene coding

Stereo scene coding [McDonald, 1993; McDonald, 1994] is an extension to temporal modulation. It avoids the problem of calibrating the projection device that impedes most active vision techniques that employ a projector and single camera from deriving accurate 3D measurements. This problem arises because it is extremely difficult to perform with a projection device the equivalent of camera calibration and space resection that require known world co-ordinate points to be associated with their detected image points. Stereo scene coding avoids this problem by employing a stereo pair of cameras that are calibrated using one of the previously described methods for stereo camera calibration.

A projection device is situated between the stereo cameras and used to project onto the scene a sequence of stripe patterns identical to that used in temporal modulation. The projected stripe pattern is reconstructed separately for each of the sequences of images acquired by the stereo pair of cameras. A map of the disparity between the

projected pattern and each image is then determined automatically exactly as it was in single camera temporal modulation. However the benefit of the stereo scene coding technique is that the two disparity maps, i.e. projector-to-left-camera and projector-to-right-camera, can be combined to derive a single disparity map that identifies correspondences between the left and right cameras. This left-camera-to-right-camera disparity map can then be used in space intersection to derive 3D measurements in the usual way for a stereo camera system.

Although stereo scene coding offers the potential for accurate acquisition of 3D measurements using active vision, it still suffers from all of the other problems associated with temporal modulation. Namely: the long capture time associated with the requirement to project a sequence of patterns and capture multiple images means that accurate measurements can only be obtained for inanimate objects, the depth resolution of the technique is limited by the minimum width of stripe that can be projected and detected, and the reliability with which the stripe pattern can be extracted is reduced in the presence of strong intrinsic texture on the imaged surface.

2.4.4. Combined active and passive methods

It is widely recognised that the success of both passive area-based and passive feature-based methods of solving the stereo correspondence problem is highly dependant on the contents of the observed scene [Cochran, 1989; Hoff, 1989; Mohan, 1989]. The number of match points that are derived by feature-based techniques is directly dependant on the number of features detected in the observed scene. While inadequate image texture in the stereo images input to area-based techniques tends to result in disparity maps that are inaccurate, noisy or incomplete. Real world scenes frequently contain neither sufficient amounts of image texture nor the necessary number of image features to allow either passive stereo matching technique to recover dense, continuous, surface maps.

In order to address this problem, several researchers have investigated projecting special illumination patterns onto the scene to artificially boost either the number of matching features or the amount of image texture in the observed scene. In this way, the performance of passive stereo matchers can be enhanced for images that would otherwise yield poor results. This technique of projecting patterned light in order to overcome the problem of insufficient scene texture has been practised in close range photogrammetry since as early as 1957 [Hertzberg, 1957].

More recently, Nishihara [Nishihara, 1987] has reported that improved results can be obtained from an adaptation of the MPG stereo algorithm when "unstructured

illumination" in the form of a binary random texture pattern is projected onto the scene. The author has also shown independently that the high spatial frequency components of images are boosted when they are captured while the scene is bathed in textured light. This has in turn been shown to result in dramatic improvements in the disparity maps derived from the MSSM stereo matcher [Urquhart, 1989; Urquhart 1990]. The term *texture enhanced stereo* has been coined to refer to this combination of active textured scene illumination and passive area based stereo matching [Siebert, 1990].

2.5. Videometrics

It can be seen from the foregoing discussion that research into the separate problems of stereo correspondence and video camera calibration is now at an advanced stage. The goal of videometrics is to integrate solutions to both of these problems into a single computer vision system that can "accurately and automatically recover 3D properties, such as co-ordinates and dimensions, from 2D intensity images" [El-Hakim, 1993]. However very few researchers have so far successfully managed to accomplish this. Most systems that employ advanced methods for accurate video camera calibration are limited by computing correspondences only for a sparse set of points denoted either by special targets markers that are physically placed into the scene or by spots of laser light. By contrast, those systems that employ robust methods for automatically determining a dense set of stereo correspondences, have largely failed to achieve levels of 3D measurement accuracy that are of widespread practical use.

A comparison of three systems that represent the state-of-the-art in videometrics illustrates the dichotomy in the performance of current systems. Aliverti, Ferrigno and Pedotti [Aliverti, 1993] have developed a system comprising four CCD video cameras that can achieve extremely high accuracy: 95% of measurements accurate to within ±0.07mm for a viewing volume of 1.7m across the diagonal. However their system calculates 3D measurements only for a single spot of laser light on the surface at a time. Therefore a dense set of measurements can only be obtained by scanning the laser across the surface one point at a time. The system takes 180s to acquire a typical surface model comprising 13000 points (i.e. less than one hundred 3D measurements per second).

Cyberware Inc. have developed a range of commercial 3D measurement systems. These use a more sophisticated laser striping system to scan a line of laser light across the surface to be measured. They can therefore acquire a dense set of surface

measurements much more quickly than a single spot scanner. For example, the latest Cyberware whole-body scanner can produce 60,000 3D measurements per second using four separate laser striping heads [Cyberware, 1995]. (Nevertheless, it should be noted that the acquisition speed of all such laser based systems is limited by the requirement to scan the laser spot or stripe across the surface). Although no figures are given in the Cyberware product information regarding the accuracy of their system, it appears that an accuracy of around ±1.0mm is possible over a cylindrical volume of height 2.0m and diameter 1.2m. Therefore the increase in speed over that achieved by Aliverti's spot laser system would appear to have been at the expense of much poorer surface measurement accuracy. It should also be noted that the Cyberware system is mechanically complex and extremely expensive.

In [Jansa, 1993], Jansa, Hang and Trinder describe a stereo scene coding system that they have used for the quantitative measurement of human bodies. This system perhaps strikes the best balance between acquisition speed and measurement accuracy of reported videometric systems. The acquisition time is quicker than the laser scanning systems, requiring only that separate images be captured when each of a series of up to eight binary patterns are projected onto the scene. However, arguably, this is still too slow for many applications. The reported accuracy of the system is better than ±1.0mm for a viewing volume of 600mm × 600mm × 600mm and mean object distance of approximately 2.25m.

2.6. Dynamic computer vision

2.6.1. The dynamic vision paradigm

During the mid to late 1980s several researchers, most notably Bajcsy [Bajcsy, 1985], Aloimonos [Aloimonos, 1987] and Ballard [Ballard, 1991], began to question the general acceptance of the computer vision community that perception is a purely passive activity involving the static analysis of passively sampled data. Instead they argued that perception is active and that perceptual activities should be capable of responding to physical stimuli. The implication of this paradigm for computer vision systems is that they should be capable of dynamically acting in accordance with information recovered from the viewed scene. This new paradigm for computer vision will be referred to as *dynamic vision* for the remainder of this thesis. One specific example of dynamic vision is the equipping of computer vision systems with some form of attention directing or gaze control mechanism so that they can actively look rather than just passively see.

2.6.2. Benefits of dynamic vision

Aloimonos, Weiss and Bandyopadhay [Aloimonos, 1987] showed that, in theory, a dynamic observer can solve many basic vision problems more accurately and more efficiently than a static one. The problems of deriving shape-from-shading, shape-from-contour, shape-from-texture and structure-from-motion are all ill-posed for a static observer and therefore it is often necessary to make unreasonable assumptions in order to simplify these problems sufficiently that they can be solved by computer. However all of these problems are shown in [Aloimonos, 1987] to become better posed and therefore may be solved using many fewer assumptions and hence more accurately when the observer is dynamic.

Some practicable benefits of dynamic vision are more obvious than the more theoretical ones explored by Aloimonos. For example, the ability to direct the gaze of a vision system allows a large area to be imaged at high resolution even when the sensor itself is relatively small. This principle is demonstrated most strikingly by the human vision system, which employs a central high resolution fovea that covers less than 0.01% of the full visual field. However our eyes continuously and unconsciously make small saccadic movements that give us the perception of a stable, high resolution view of a much larger proportion of the visual field [Ballard, 1991]. Ballard notes that a key feature of such a dynamic vision scheme is an intelligent controller that can adapt the gaze directing strategy both to the particular task that is being undertaken and to the contents of the observed scene.

The integration of gaze control and attention directing mechanisms into a stereo computer vision system provides additional practical benefits. It permits piece-wise surface reconstruction from scenes that would otherwise exceed the field of view or fusional limits [Das, 1990] of a static system. It also supports the use of a verging stereo vision system, i.e. one in which both cameras can be fixated onto the same point in the scene. This ensures the maximum amount of overlap between captured stereo images, resulting in bipolar disparities and reducing the search required to establish stereo correspondences to a minimum. A verging stereo system can also accommodate very efficient foveated vision strategies such as that described by Ballard [Ballard, 1991].

2.6.3. Dynamic vision heads

Some initial examples of a new generation of stereo computer vision systems that comprise steerable cameras to provide support for the dynamic vision paradigm have been reported. This class of dynamic stereo vision platforms has become known as

heads because of their obvious analogy to the human oculomotor system. One of the first, and also perhaps one of the most successful, of these heads has been the one developed at the University of Pennsylvania. This was first proposed by Bajcsy [Bajcsy, 1985] and actually developed by Krotkov [Krotkov, 1989].

The Pennsylvania head has a total of eleven degrees of freedom that are actuated under computer control. Two CCD video cameras are mounted on coupled pivots in such a way that they can be converged or diverged by a single motorised screw. The camera pivots are separated by a fixed distance of 128mm, so this figure is the nominal stereo baseline for the system. The whole vergence platform can be rotated in azimuth and elevation by two separate A.C. servomotors, while two further A.C. servomotors provide controlled translation in the X-Y plane. A separate wire-wound potentiometer senses the position of each of these four A.C. servomotors. The zoom, focus and aperture of both cameras are controlled by a total of six variable speed D.C. servomotors that are coupled through optical shaft encoders. Further flexibility is provided by a number of light sources whose intensity of illumination can also be controlled by computer.

Krotkov [Krotkov, 1989] implemented two methods for deriving quantitative depth information that rely greatly upon the dynamic capabilities of the Pennsylvania head. The first method is "range from focus", that derives an estimate of the range to an object from the focus position of the lens required to bring the image of the object into sharp focus. The second method is a fairly unsophisticated stereo line matching algorithm that provides disparity estimates from captured stereo images of the scene. Range is computed from the obtained disparity estimates using a simple function of the vergence motor position. However the reported accuracy of each of these methods of range estimation is poor (approximately 1.0%/m R.M.S. range error for focus ranging and approximately 2.3%/m R.M.S. range error for stereo ranging[3]), although Krotkov reports that slightly improved accuracy can be obtained by combining statistically range data from both sources. The calibration and the accuracy of the Pennsylvania head are discussed in greater detail in sub-section 2.6.4.

Several other dynamic stereo vision systems have been reported in the literature including: the Rochester Robot [Ballard, 1988], the system at the Royal Institute of Technology in Stockholm [Pahlavan, 1991], the Turing Institute's Richard 1st

[3] By Krotkov's definition, a range error of 1.0%/m means an error at a range of 1.0m of 1.0% of 1.0m = 1.0cm, an error at a range of 2.0m of 2.0% of 2.0m = 4.0cm, and so on.

[Undbekken, 1991] and Sheffield University's COMODE system [Thacker, 1992]. Other institutions such as the University of Aalborg, Genoa DIST, LIFIA in Grenoble and the University of Surrey are also known to be developing similar dynamic stereo vision systems [Buxton, 1991]. However the main factor governing the development of all of these systems has been that they replicate as closely as possible the degrees of freedom and speed of movement of the human occulomotor system. They have therefore not been designed to permit quantitative measurement. Krotkov's system at the University of Pennsylvania is the only reported dynamic vision system that, so far, has been shown to be capable of deriving quantitative measurements from an imaged scene.

2.6.4. Calibration of dynamic vision systems

Just as a static stereo vision system must be calibrated before it can derive quantitative measurements of an observed scene, so must a dynamic stereo vision system. However, the very nature of dynamic vision results in systems whose extrinsic orientation parameters vary dramatically with every change in orientation of the cameras. Therefore the methods of calibration described in section 2.3, that mainly rely upon the static imaging of a fixed calibration grid, are not directly applicable to dynamic stereo vision systems. The relative infancy of the field of dynamic vision and the difficulty of the task of calibrating dynamic vision systems has meant that only a few techniques have been investigated so far. Those few techniques that have been investigated are still very much in their early stages of development. However two distinct approaches to this problem are beginning to emerge: those of direct calibration and self calibration. These are described below.

2.6.4.1. Direct calibration

The direct method of calibrating a dynamic stereo vision system uses feedback from the system's actuators to maintain an up-to-date estimate of the exterior, and possibly also the interior, orientation parameters. The Pennsylvania head [Krotkov, 1989] is perhaps the only reported dynamic stereo vision system to have successfully employed this method of calibration. Krotkov determined empirically the relationship between the focus motor position that brings an imaged point into sharp focus and the range to that point for the Pennsylvania head. Similarly, he also demonstrated that the disparity at an image point and the position of the vergence motor could be used to determine the range to the point. Although both of these cases are rather simplistic, each effectively calibrating only one degree of freedom of the Pennsylvania head at a time,

they do provide initial evidence of the feasibility of direct calibration of dynamic vision systems.

One criticism that has been levelled at Krotkov's attempts to calibrate the Pennsylvania head has been the poor level of ranging accuracy that was obtained. The accuracy of the approach clearly relies upon the sensor platform being highly stable mechanically and also being capable of providing precise positional feedback from its actuated components. However, in his book [Krotkov, 1989], Krotkov acknowledges that the Pennsylvania head was built at a low cost and largely from surplus equipment and spare parts. The low level of accuracy obtained is therefore due very possibly to the mechanical shortcomings of the system and does not result from an inherent deficiency in the direct calibration approach itself. The simplistic model that was used for verged imaging and the relative inaccuracy of the rather unsophisticated stereo matcher that was employed also contributed to the poor level of accuracy obtained for stereo ranging.

The investigation of Driels and Pathre [Driels, 1991] provides further support for the potential of direct calibration. They have developed a vision-based automatic theodolite (VBAT) that is essentially a monocular dynamic vision system. The VBAT system consists of a video camera that is mounted on standard off-the-shelf rotation stages driven by stepper motors that provide computer controlled rotation in azimuth and elevation. The resolution and accuracy of the azimuth rotation stage are both $\pm 0.01°$ while the elevation rotation stage has a resolution of $0.01°$ and accuracy of $\pm 0.05°$. The system is intended to track automatically a precision, internally illuminated, target sphere that is attached to the end-effector of a robot. The position of the robot end-effector is to be determined automatically from the rotation angles indicated by the rotation stage encoders and the detected location of the centroid of the image of the target sphere in the acquired video imagery. Presumably a second VBAT system will be required in practice to allow triangulation of the three-dimensional position of the illuminated target sphere.

In order to calibrate the VBAT system and to verify its accuracy, Driels and Pathre attached the target sphere to a co-ordinate measuring machine (CMM). The target sphere was moved to 90 accurately known locations using the CMM and tracked with the VBAT system. The data obtained was used to determine 4 vision system parameters and 12 kinematic parameters for the VBAT systems. Four kinematic parameters were used to represent each of three consecutive Denevit-Hartenberg links between the world, azimuth rotation stage, elevation rotation stage and camera co-ordinate frames. A further independent set of 18 test target locations throughout the

working volume of the CMM was then used to measure the accuracy obtained by the VBAT system. The results of this test indicated an R.M.S. miss distance of ±0.2mm for a target range of approximately 4.0m. (The miss distance is the perpendicular distance from the line of sight of the VBAT system to the target.)

2.6.4.2. Self calibration

None of the current crop of dynamic stereo vision systems appear to have been designed or constructed with sufficient precision to permit a practically useful level of accuracy using direct calibration. The alternative approach of self calibration as a potential solution to the problem of calibrating dynamic stereo vision systems has therefore recently begun to attract interest. The goal is to calibrate a stereo pair of cameras purely from the image locations of matched pairs of corresponding points. No *a priori* knowledge of the image scene, such as measured calibration grids, or of the camera parameters, e.g. as provided by accurate actuator feedback in the direct approach, are therefore required. Consequently, self calibration has become especially desirable to the developers of the current dynamic stereo vision systems which are, in general, poorly engineered.

One of the main contributors to the field of self calibration has been Faugeras. Together with his colleagues at INRIA, he has been investigating this problem since 1986 [Faugeras, 1986]. The basis of their solution is the epipolar constraint that restricts the possible positions for pairs of corresponding points in a given stereo imaging configuration. (In photogrammetry, the epipolar constraint is better known as the coplanarity constraint.) Their approach is to search for a combination of exterior orientation parameters that minimises the discrepancy of matched stereo pairs of points from the epipolar constraint. The exterior orientation parameters can thus be determined purely from the locations of matched pairs of points after each movement of the cameras. In the method described in [Toscani, 1989] this process must be preceded by an initial stage of conventional static camera calibration to determine the interior orientation parameters of the stereo cameras. The interior orientation parameters are assumed to remain constant thereafter.

In 1992, Maybank and Faugeras showed that the interior orientation parameters of a single moving camera can be determined by tracking the image locations of scene points in at least three captured images [Maybank, 1992]. More recently a modified method has been reported in [Zhang, 1994] that requires two sets of stereo images to be acquired with the stereo cameras in different positions. Stereo correspondences and temporal correspondences are then established within and between both stereo pairs of images respectively. This information then allows the interior orientation parameters,

excluding the position of the principal point, as well as the exterior orientation, parameters to be determined using the epipolar constraint.

Unfortunately, only low levels of accuracy have been obtained to date using self calibration methods. It is also only possible to obtain solutions for the camera parameters up to a certain scale factor. These two factors have therefore prevented the use of self calibration methods in quantitative tasks. A further disadvantage of the approach is the complexity of the problem and the large number of degrees of freedom of the search space. This requires iterative solutions that are often difficult to bring to convergence without additional constraints. This contributes significantly to the poor levels of accuracy that have been reported so far.

Many similarities are evident between the self calibration approaches proposed by the computer vision community and photogrammetric methods such as self calibrating bundle adjustment. However it appears that the photogrammetric methods employ statistically based analyses that are more rigorous than those proposed by the computer vision community. It may well be that much better accuracy could therefore be obtained by applying photogrammetric methods to the problem of calibrating dynamic computer vision systems.

2.7. Conclusions

The literature reviewed in this chapter indicates that the development of a fixed viewpoint stereo 3D computer vision system is a challenging problem that is still some way from being "solved". This problem is therefore clearly worthy of investigation in its own right. However, by choosing to develop a stereo 3D computer vision system, such as the proposed ASP stereo vision system that employs steerable cameras, it is also clear that the problem is made even more challenging.

In order to be successful, dynamic and static stereo vision systems alike must address the two basic problems of stereo correspondence and system calibration. It is apparent from the literature that the problem of calibrating a fixed viewpoint stereo vision system is becoming increasingly well understood and that mature algorithms for its solution are now becoming available, e.g. [Tsai, 1987]. Although research into computer stereo correspondence is also well established, no completely satisfactory solutions have yet been developed. It is the inadequacy of the current state-of-the-art in solutions to the stereo correspondence problem that is most severely hindering the development of robust and broadly applicable fixed viewpoint stereo vision systems.

Although dynamic computer stereo vision has previously attracted much less interest than fixed-viewpoint stereo vision, a small number of examples of this type of system have been reported in the computer vision literature, most notably [Ballard, 1988; Krotkov, 1989; Pahlavan 1991]. These systems have proven to be useful test beds for researchers investigating the basic principals of dynamic vision: vergence, saccadic gaze control and smooth pursuit. However, almost none of the papers describing dynamic computer stereo vision systems have reported their successful use in any practical application, except perhaps as remote sensor heads for "tele-presence" tasks. Surprisingly, few appear to have successfully utilised the basic benefit, indeed purpose, of employing two cameras: the ability to judge depth or measure distance automatically.

The inability of previously reported dynamic stereo computer vision systems to provide quantitative measurements with a useful level of accuracy can be largely attributed to the lack of calibration techniques suitable for dynamic systems. Those techniques that have been successfully developed for calibrating static systems are not readily applicable to dynamic systems with steerable sensors. In addition, stereo correspondence is equally important to dynamic stereo vision systems that are to perform 3D measurement as it is to static systems. The inability of previously reported dynamic stereo vision systems to perform accurate 3D measurement can therefore also be partly attributed to the inadequacy of the current state-of-the-art in solutions to the stereo correspondence problem.

Chapter 3
Approach

3.1. Objectives

The work described in this thesis comprised part of the Active Stereo Probe project. The goal of the Active Stereo Probe project was to develop, construct and test a generic computer vision system capable of dynamically probing and interacting with its environment. The basic function of the ASP system was to be that of a 3D stereo computer vision system. However its functionality was to be considerably enhanced over that of existing systems through the combined use of a novel active projection source and steerable sensors. Furthermore, it was proposed that a range of low-level automatic reflexes (such as head orientation, camera focus, vergence and gaze control) and an "intelligent" high-level control system would be integrated into the ASP system. This would allow it to function either semi-autonomously or even fully autonomously for some applications.

The final ASP system was planned to provide a range of low-level visual measurement and analysis functions such as automatic vergence and tracking of objects in 3D space, generation of 3D surface models of objects and object metrology and recognition based on interaction with the imaged scene. These functions were to be combined together under the control of the "intelligent" high-level controller. The high-level controller would then be tuned to allow the combined system to be used in a wide range of advanced applications in areas as diverse as medicine, autonomous robot guidance, virtual reality and surveillance. However, it should be noted that the investigations into low-level visual reflexes and "intelligent" high-level control were the responsibility of other members of the ASP project team and are therefore not discussed in this thesis.

The prime objective of the work presented in this thesis was to develop the underlying stereo 3D computer vision component of the ASP system. This comprises three intimately connected modules: a *sensor head module* consisting of the actuated binocular cameras with their associated control hardware and software, a *stereo*

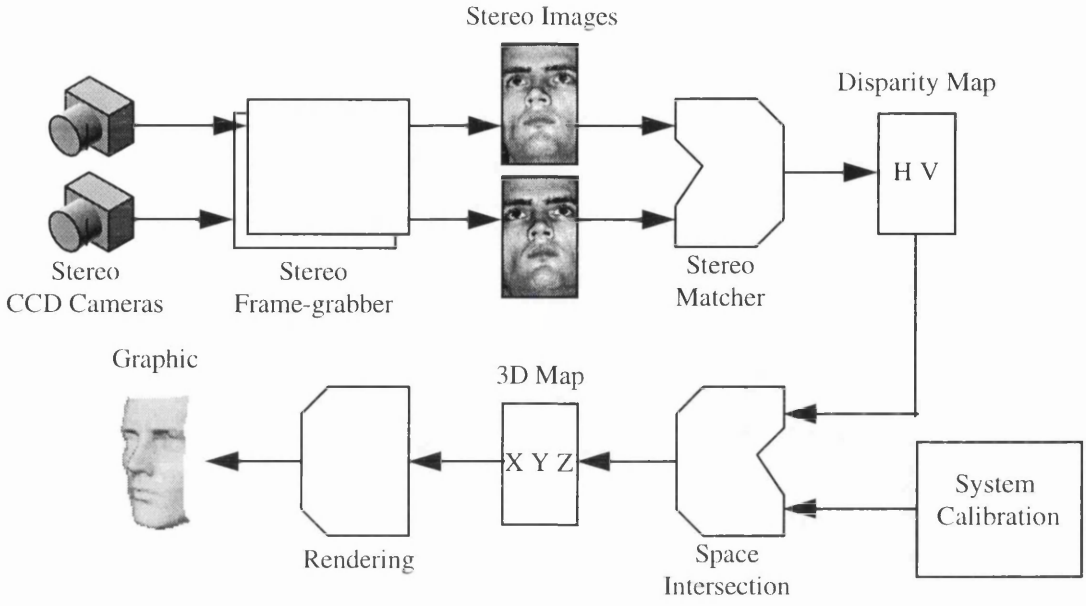


Figure 3.1. Basic components of a stereo 3D computer vision system.

matching module and a *system calibration module*. These modules are described in detail later in this chapter; however, the basic process by which they combine to recover 3D measurements from an object placed in the field of view of the stereo cameras operates as follows:

1. A stereo pair of images of the object is acquired from the CCD cameras by the frame-grabber.

2. The acquired images are processed by the stereo matcher to calculate the image correspondences between the left and right images. This information is output in the form of a *disparity map* which contains the horizontal (H) and vertical (V) offsets between the image locations of each pixel in the left image and the calculated matching point in the right image.

3. Space intersection is then performed on the disparity map to recover the 3D surface of the imaged object. This process requires that system calibration has first been carried out.

Although it was not a direct requirement for the ASP system, it was decided that a second objective of this work would be to integrate the above three modules with a fourth, *3D visualisation* module, as illustrated in figure 3.1. This would result in a complete stand-alone stereo 3D computer vision system that would serve two purposes. Firstly, it would allow the 3D vision components of the ASP system to be

tested and demonstrated (while under the control of a human operator) prior to the completion of the full ASP system. (The implementation of the low-level automatic reflexes and the "intelligent" high-level controller to provide the full ASP system with the capability for autonomous operation were outwith the scope of this work). Secondly, it was expected that this stand-alone stereo 3D computer vision system could, by itself, be used in a wide range of applications.

3.2. Review of problem

3.2.1. Overview

The literature review of Chapter 2 has shown that the two problems that have primarily been associated with the development of a dynamic stereo computer vision system are those of computer-based stereo correspondence and dynamic calibration. These problems are discussed in detail in sub-sections 3.2.3 and 3.2.4 following. However there is an additional problem that is perhaps just as important to the development of a dynamic computer stereo vision system but that has been largely overlooked in the literature: that of the design of the sensor system itself. This problem is discussed first in sub-section 3.2.2.

3.2.2. Sensor system design

A rigorous scientific method for selecting the fundamental parameters of a stereo vision system (namely: the camera resolution, the focal length of the camera lenses, the baseline separation of the cameras, the camera to object distance and the working volume) has not been reported in the computer vision literature. It can only be deduced that the majority of previous stereo vision systems have not been rigorously designed at all and at best they have been constructed using a mixture of heuristics and experimental trial and error. Dynamic systems in particular appear to have been largely assembled from spare components available in the laboratory.

The lack of a rigorous design methodology is reflected in the haphazard approach that has until now been adopted for the design of computer stereo vision systems in general, and dynamic systems in particular. It is argued in this thesis that this "let's just build something and then see what we can do with it" approach has contributed greatly to the failure of dynamic stereo vision systems to achieve accurate results. The fundamental parameters listed above establish the basic performance capabilities of a stereo vision system. It is therefore argued that a desired level of accuracy is only likely to be obtained if it is specified at the earliest possible stage of development and

then the fundamental parameters of the system are chosen in order to meet this specification. The adoption of such an approach has not previously been reported. It is therefore clear that a new methodology for designing dynamic computer stereo vision systems must be developed. This topic is explored in detail in Chapters 4, 5, 6 and 7.

3.2.3. Stereo correspondence

A fundamental component of any computer stereo vision system is a computer algorithm capable of automatically determining correspondences between a stereo pair of images. The three basic requirements of such an algorithm are:

- *Accuracy*: the stereo vision system can only provide accurate 3D measurements if the stereo correspondences, or disparity measurements, on which the 3D measurements are based are accurately determined.

- *Speed*: a stereo correspondence algorithm that requires several hours or even days of processing is unlikely to be useful in any practical application. In a dynamic stereo vision system, real time performance (i.e. rates of around 25Hz) is the ultimate goal and execution times of the order of a few minutes are probably the absolute limit of acceptability.

- *Density*: in the majority of applications, a dense set of surface measurements, and hence stereo correspondences, are required to accurately represent the surface that is to be measured. This is particularly true of many applications where the surface is not a simple geometric shape that can be interpolated from a small number of measurements (e.g. the surface of a patient's body in medical applications).

Although the research of stereo correspondence algorithms has received much attention, no examples of algorithms that meet all three of the above requirements have been reported. The research has followed two main lines of investigation. The first is into feature based stereo matchers that attempt to find correspondences between features, such as corners, that are detected in both of the stereo images. Such algorithms have been reported to achieve close to real-time performance using specialised hardware. However they are inherently unable to provide a dense set of measurements because of the limited number of the required features that appear in imaged scenes and that can thereafter be individually identified and processed. This seems to limit their use to laboratory conditions and a few specialised industrial applications where the measured surfaces are easily interpolated geometric shapes.

The second, and more promising, class of stereo correspondence algorithms that have been researched are correlation-based stereo matchers. These are capable of providing a dense set of stereo correspondences: one for each pixel in the input stereo images. The most sophisticated of these algorithms can also provide sub-pixel disparity measurements and are therefore suitable for high precision surface measurement. However, the enduring drawback of this class of algorithm has been their high computational content and hence long processing times.

Despite representing the state-of-the-art in correlation-based stereo matchers, the MSSM stereo matching algorithm that was available for use at the start of the ASP project still suffered from the two shortcomings that were described in Chapter 1. The first was the perennial problem with this class of algorithm - that of slow execution: several hours of processing were required for a stereo pair of 256×256 pixel images. The second shortcoming was that no accuracy figures had ever been compiled for the algorithm. Therefore, although the disparity maps produced by MSSM appeared to be of a high quality, particularly when textured illumination was used during image capture, and the theoretical limits of the algorithm were known, its actual performance characteristics were all but unknown. Chapters 8 and 9 describe how these shortcomings of the MSSM stereo-matcher were addressed.

3.2.4. Dynamic calibration

A large amount of work concerning the calibration of static stereo vision systems has been reported in the computer vision literature. However these static methods rely almost exclusively upon special targets, whose position has previously been measured, being viewed by the cameras. This is simply unacceptable for the calibration of a dynamic system as it would require either that a special calibration object be moved to each new viewpoint or that an adequate number of target points be provided at all possible viewpoints. Thus, not only would the flexibility of the system be seriously undermined but its practicality in terms of the dynamic vision paradigm would also be compromised.

The calibration of dynamic stereo vision systems has attracted considerably less attention than that of static systems. The main thrust of the research that is now being carried out in this area is to develop a completely general method of calibration using only direct observation of objects already in the natural scene. Although the objective of achieving this "holy grail" in calibration is laudable, it currently seems far from fruition. Furthermore, the methods currently under investigation are reported to rely upon complex iterative algorithms that impose a high computational burden. This is at

odds with the requirement of dynamic vision for fast control loops and the ultimate goal of real-time operation.

An alternative approach to dynamic calibration that can potentially achieve very fast execution rates has been investigated by a few researchers. This approach uses the simple concept of determining the position and orientation of the cameras from direct positional feedback from the camera actuators. Unfortunately the accuracy that has been achieved using such an approach has been disappointingly poor. It appears that this can be attributed largely to the mechanical inadequacy of the sensor systems that have been employed. The inadequacy of the sensor systems can in turn be attributed to the "let's just build something and then see what we can do with it" approach that has been used in their development. A more rigorous scientific approach to sensor system development has not previously been possible because the relationship between the mechanical tolerances of the sensor system and the achievable accuracy of derived 3D measurements was unknown.

Note that the work described in this thesis does not attempt to perform continuous calibration of the dynamic stereo vision system while the sensors are in motion. Instead it is assumed that sensor motion will be in a series of start-stop movements, with calibration being performed only when the sensors are stationary.

3.3. Approach

3.3.1. Overview

The widespread failure of dynamic stereo computer vision systems to exploit their inherent depth measurement capability can be attributed largely to a lack of early consideration for the calibration requirements of quantitative measurement. Indeed, few dynamic vision researchers seem to have considered these requirements at all. Those few who have done so have only taken them into consideration after the construction of their dynamic binocular sensor system [Krotkov, 1989]. This has proven to be at too late a stage in their development, since the constructed sensor system is inherently too unstable or inaccurate to be calibrated accurately using the techniques that are currently available. The root of this problem is the "let's just build something and then see what we can do with it" approach that has become the norm for the development of dynamic binocular sensor heads.

It was therefore apparent that a radically different development approach was required if dynamic vision systems were to become suitable for practical, quantitative, applications. Accordingly, one objective of this work was to develop the ASP stereo

vision system using a novel approach that would consider the requirements of quantitative computer vision from the earliest stages of the system's design. It was intended that this would result in a sensor system that was capable of providing 3D measurements of a desired level of accuracy using a calibration scheme based upon direct positional feedback.

3.3.2. Performance specification

The starting point for the approach adopted for the development of the ASP stereo vision system was the specification of its basic performance requirements. These are described in section 3.4. By considering these overall requirements during the initial stages of the system's development, it was possible to identify which requirements were relevant to each of the ASP stereo vision system's individual modules. Each module could then be designed and its performance characterised as necessary to ensure that it met its own individual required level of performance. When all of the modules were then integrated to produce the completed system, this would minimise the technical risk of failing to meet the specified overall performance requirements. Accordingly, it was expected that the final ASP stereo vision system would be capable of providing a useful level of performance in practical applications.

3.3.3. Sensor system module development

There were two main problems to be addressed in the development of the sensor system module of the ASP stereo vision system:

3.3.3.1. Sensor system configuration

The first problem to be addressed was the selection of the main imaging components of the sensor head, e.g. the cameras and frame-store, and the determination of their geometrical and optical arrangement, i.e. the camera baseline, focal length of the camera lenses and the vergence angles. It was essential that the imaging components be selected as early as possible since they would have a fundamental influence on every other aspect of the development of the sensor head.

Having chosen the basic imaging components, the principal consideration in determining their geometrical and optical arrangement was the requirement to meet the working volume and disparity limit specifications described in section 3.4. However, the relationship between these factors and the parameters to be determined, camera baseline, focal length and vergence angle, was found to be too complex to be derived from a direct systems analysis alone. Consequently, it was decided that a

model of the binocular sensor head was necessary to allow problem space to be explored more easily.

A mathematical model of the sensor head was therefore derived and simulation software was developed to drive it. This model and its software implementation are described in Chapter 4. Using the completed model, it was indeed possible to analyse the relationship between the achievable working volume and disparity limits and camera baseline, focal length and vergence angle. From the results of this analysis, it was possible to predict a combination of these parameters that would satisfactorily meet the specified performance requirements. This analysis and the obtained results are detailed in Chapter 5.

3.3.3.2. Sensor system actuation

The second problem to be addressed in the development of the sensor head module was the selection of the camera actuators and the specification of construction tolerances for their supporting rig. The main parameters to be determined were the required level of encoder accuracy and the maximum acceptable level of actuator wobble and translational inaccuracies. It had been assumed that these parameters would critically influence the achievable level of stereo measurement accuracy. However, the relationship between the parameters and measurement accuracy was also found to be somewhat complex.

Therefore, it was again decided to use the sensor head model to aid the required analysis. This time the model was used to simulate the effects of encoder accuracy, actuator wobble and translational inaccuracies on stereo measurement accuracy. The resulting series of simulation experiments established the predicted levels of component accuracy and tolerance that would be necessary to meet the specified stereo measurement accuracy requirements. These experiments are detailed in Chapter 6.

3.3.3.3. Sensor system fabrication

Using the approach described above, it was intended that the performance of the sensor head module of the ASP would be extensively characterised prior to its construction. Consequently, the process of selecting suitable components and specifying construction tolerances would be greatly simplified. This would, in turn, considerably enhance the possibility of the completed system meeting the specified requirements and thereby achieving a practically useful level of performance. Chapter

7 describes the selection of the components for, and the fabrication of, the ASP sensor system.

3.3.4. Stereo matching module development

The performance requirements that would be most strongly influenced by the stereo matching module of the ASP stereo vision system were the speed of surface model generation, the measurement accuracy, and the disparity limits.

The speed of stereo matching is clearly a critical factor in the time taken to produce a surface model. Unfortunately, the MSSM stereo matcher that was made available for use in the ASP project was slower than the specified requirements by several orders of magnitude. Considerable effort, both by the author and others, most notably Arthur van Hoff, was therefore spent in re-implementing the MSSM algorithm to improve its speed. Details of the various stages in the development of the MSSM stereo matcher are given in Chapter 8.

One of the express purposes of the simulation analysis mentioned in sub-section 3.3.3 was to choose an imaging configuration that would accommodate the disparity limits imposed by MSSM. Therefore, there was no specific requirement to alter this characteristic of the MSSM algorithm.

The theoretical matching accuracy provided by the MSSM stereo matcher was well within that required to meet the measurement accuracy specification for the ASP stereo vision system. However, the actual matching accuracy achieved in practice by the original version of the algorithm was largely unknown. This was primarily because the slow speed of this version prohibited the large amount of testing needed to fully characterise its performance. In any case, there appeared to be little to be gained from characterising the performance of an old version of MSSM when a new version would be used in the final ASP system. An immediate priority following the development of the new, and expectedly much faster, version of MSSM was therefore to thoroughly characterise its performance to ensure that it was capable of meeting the ASP performance requirements specifications.

3.3.5. System calibration module development

A stereo vision system measures a 3D surface by forming the image of a number of points from the surface on the image planes of the two cameras. After stereo matching, the 3D position of each of the imaged surface points is then recovered (or triangulated) by back-projecting its stereo pair of image points using some form of

model of the stereo vision system. In common with most reported stereo vision systems, the collinearity equations (see section 2.2.2) are to form the basis of this model in the ASP stereo vision system.

The purpose of the system calibration module of the ASP stereo vision system is to determine the parameters of the collinearity equation model. Accurate 3D measurement requires that these parameters be determined accurately since surface measurement error, the error between the "actual" and "measured" positions of the surface points, will be introduced through any errors in the collinearity equation model. (Surface measurement error will also be introduced through errors in the measured position of the stereo pair of image points).

In developing the calibration module for the dynamic ASP stereo vision system, the main problem to be addressed is the requirement to maintain calibration of the stereo cameras after they have moved. The objective was to devise and investigate a direct approach to dynamic stereo camera calibration that uses positional feedback from the sensor actuators to maintain an open loop estimate of sensor position and orientation. Thus, after each sensor movement, the external orientation parameters in the collinearity equation model could be updated appropriately using the newly obtained sensor positions.

Although only very limited accuracy has been previously reported using this approach, it was expected that the extensive system modelling employed in the design of the ASP system would ensure that much greater accuracy would be achievable. A novel extension to this approach that was expected to achieve greater accuracy was then to close the calibration loop by using information provided from stereo correspondences in images captured at the new viewpoint. A three stage approach to system calibration was therefore devised and adopted for further investigation. Each of the stages of the approach are described in more detail in sub-sections 3.3.5.1, 3.3.5.2 and 3.3.5.3.

3.3.5.1. Static calibration

Static calibration is the first stage in the proposed dynamic calibration cycle. It is carried out "off-line", prior to dynamic operation of the sensor system, while the sensors are held in a fixed position. The standard technique used in the calibration of static 3D vision systems of imaging a measured target grid is perfectly adequate for this stage of calibration. Static calibration obtains both the internal orientation parameters (e.g. the focal length and image centre position) of the collinearity equation model that will not alter during dynamic operation and initial estimates for the external orientation parameters (the position and orientation of the cameras) that

will vary during dynamic orientation and that therefore must subsequently be updated using dynamic calibration. Chapter 10 describes in more detail the development of the static calibration stage.

3.3.5.2. Open-loop dynamic calibration

The second stage in the proposed dynamic calibration cycle is *open-loop dynamic calibration* which is carried out immediately after each sensor movement. This stage has been termed open-loop, because (in contrast to the third stage of *closed-loop dynamic calibration* described below) it does not use any image derived information to close the calibration loop. Instead, the current estimate of the external orientation parameters, derived either from static calibration or from the previous round of dynamic calibration, is updated using direct encoder feedback from the sensor actuators. The advantage of this approach to calibration is that its direct nature requires the minimum of computer processing and does not require the use of any form of targets in the viewed scene. Therefore, it is eminently suitable for use in a dynamic vision context.

However the success of this approach is clearly dependant on the accuracy of the positional feedback and the overall stability of the sensor head rig. The few reported attempts at utilising this approach to dynamic calibration [Krotkov, 1989] have failed precisely because the sensor heads that were employed had not been well enough engineered and the actuator feedback lacked sufficient accuracy. It appears that these deficiencies were due to the required levels of accuracy and stability not being analysed sufficiently deeply, or indeed not being analysed at all, prior to the construction of the sensor head. In contrast, the approach that has been adopted in this work attempts to address this problem by performing the rigorous pre-construction system modelling and characterisation analysis that is outlined in sub-section 3.3.3 and detailed in Chapters 4, 5 and 6.

3.3.5.3. Closed-loop dynamic calibration

A third stage of *closed-loop dynamic calibration* is also proposed in the dynamic calibration cycle. As its name implies, the objective of this stage is to close the dynamic calibration loop using information derived from captured images. This would allow both the external orientation parameters derived from open-loop calibration to be refined for greater accuracy and the internal orientation parameters to be tracked in the event of drift.

It was decided to investigate the use of the powerful photogrammetric technique of *bundle adjustment*, e.g. [Brown, 1971; Faig 1975; Wong 1975; Granshaw, 1980] within the closed-loop dynamic calibration process. The use of this technique in a computer vision system such as the ASP system had not been previously reported. Accordingly, an implementation of bundle adjustment, tailored for the specific requirements of the ASP stereo vision system was needed. Its main purpose was to use the image co-ordinates of stereo pairs of image points, matched using the MSSM algorithm, to derive new estimates for the internal and external orientation parameters in the collinearity equation model. An initial investigation into the implementation of a bundle adjustment program is presented in Appendix A.

However, in contrast with the requirements of dynamic vision, bundle adjustment is generally a slow process requiring many iterations to derive an accurate solution. A novel method to reduce the required number of iterations and achieve a more appropriate level of performance was therefore proposed. The proposal was to use estimates of the external orientation parameters derived from open-loop dynamic calibration, i.e. from encoder feedback, to reduce the search space required by the bundle adjustment. It was hoped that this would result in much quicker convergence to an accurate solution. Hence both the high accuracy and high speed required by the calibration module of the ASP stereo vision system could be achieved.

3.4. Performance requirements

An important initial stage in the development of any system is the identification and specification of the performance requirements for that system. However, initially it was not possible to specify in detail the requirements for the ASP stereo vision system since the realisable levels of performance for such a system were largely unknown. Indeed, the primary objective of many of the experiments described in the coming chapters was to explore the limits of achievable performance.

Nevertheless, it was possible to define the following specification of the basic level of performance that would be acceptable for the ASP stereo vision system. This specification was based upon the experience that had been obtained from earlier studies and both the capabilities of the MSSM stereo-matcher and the basic requirements of the application domain chosen for particular attention: human body metrology or *biostereometrics*.

3.4.1. Measurement accuracy

The most important aspect of the requirements specification for any measurement system is the accuracy with which it can measure. The practical limits of measurement accuracy for the ASP stereo vision system were again largely unknown. However, in order to be of any practical use in the chosen application area, the system would have to be capable of recovering 3D surfaces to an accuracy of within the order of a few millimetres. A figure of ±5mm error in each axis was therefore adopted as the minimum acceptable 3D measurement accuracy.

3.4.2. Working volume

It is an obvious requirement of a vision system intended for human body metrology that it must be capable of viewing the entire volume of a human body. However it is also a competing requirement of a stereo vision system that the surface to be measured should be viewed at as high a resolution as possible. It was clear from previous work that stereo video images of an entire body would be at far too low a resolution to allow surface recovery with a useful level of accuracy.

A compromise solution was therefore adopted that would utilise the steerable viewpoint capability of the ASP system. This solution specified that the *viewing volume*, i.e. the volume of the scene visible by both cameras at any instant, should be a cuboid of maximum dimensions 1.0m high × 1.0m wide × 0.5m deep. It was intended that this viewing volume would be sufficiently large to enclose most parts of the human body that may be of interest to a clinician (e.g. part of a limb, a hand, a foot, the whole of the head or the chest), but would also be small enough to permit video imagery of sufficiently high resolution.

In the event that a larger volume is required, this could be achieved by scanning the actuated cameras and hence the vergence or fixation point at the centre of the viewing volume to different locations, thus providing an extended *working volume*. By specifying that the vergence point could be moved to any point within a cube of dimensions 1m × 1m × 1m, a working volume of 2m wide × 2m high × 1.5m deep could be achieved. This working volume, which is illustrated in figure 3.2, was deemed large enough to enclose the vast majority of human bodies and also to allow the viewpoint to be adjusted as required.

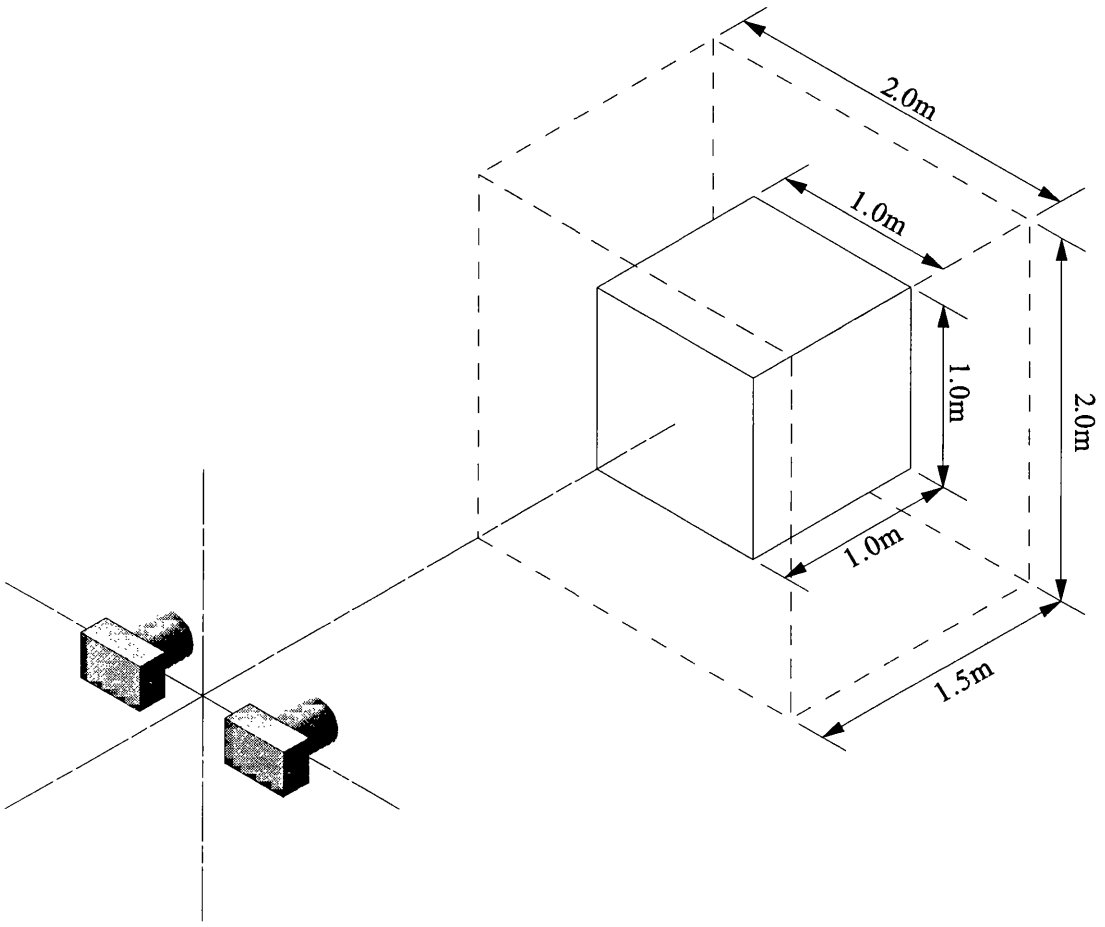


Figure 3.2. Preliminary specification of the working volume (dashed lines) and volume of possible vergence points (solid lines) for the ASP stereo vision system.

3.4.3. Disparity limit

In order that a stereo 3D vision system be capable of deriving a measurement of the position of a given point it is not sufficient that the point be visible to both of the cameras. It is also necessary that the disparity obtained by stereo observation of the point be within the *disparity limit* of the stereo matching algorithm. Hence the depth of the viewing volume, described in the previous section, may be constrained by the disparity limit of the MSSM stereo matching algorithm that was adopted for use in the ASP stereo vision system.

The practical upper limit of disparity for the version of the MSSM software that was available at the time when the Active Stereo Probe was being designed was dictated by the length of time required to process stereograms containing large disparities. Essentially the required processing time increased as the square of the maximum

disparity in the stereogram to be matched. Hence it was desirable to limit as much as possible the maximum permissible disparity.

Furthermore, in order to recover the small vertical component of disparity that results from a verged stereo imaging geometry, a two dimensional search is required in the stereo matching process. This two dimensional search introduced a very considerable processing overhead into the version of MSSM available at that time. It was therefore also desirable to limit the amount of vertical disparity in stereograms captured by the Active Stereo Probe, while maintaining a verged stereo imaging geometry.

However, if the range of disparities present in a stereogram of a scene is too small, the result is an associated loss of dynamic range of disparity and hence ranging resolution. Consequently, in addition to the upper limit on the range of disparities imposed by processing overheads, there is also a lower limit imposed by the required ranging resolution. An upper limit of ±20pixels and a lower limit of ±10pixels was therefore chosen as the basic specification for the allowable range of horizontal disparities. Since the MSSM algorithm was believed to have a sub-pixel matching precision of $^1/_{15}$ of a pixel, this choice of horizontal disparity limit would provide a range of between 150 and 300 quantised measurement steps. An upper limit of ±0.5 pixels was chosen as the preliminary specification for vertical disparity so that the search process could be constrained as closely as possible to one dimension.

3.4.4. Measurement speed

Another important performance factor that had to be specified for the ASP stereo vision system was the speed with which it could make surface measurements. The ideal level of performance, and the long term goal of 3D stereo vision research, is the production of complete, dense surface measurements in real-time, i.e. at video frame rates. This equates to the production of millions of point measurements per second: a level of performance that is still far from being achievable.

Fortunately, for many applications, including human body metrology, such high measurement rates are not required. It was therefore deemed sufficient that the ASP stereo vision system be capable of the on-line production of complete surface models. i.e. within the order of a minute or two. This equates to a more realistically achievable figure of around 5000 point measurements per second.

3.4.5. Degrees of freedom

The basic requirement for camera actuation was that the cameras should be capable of scanning the extended working volume described above and providing suitably verged stereo images. Hence the following degrees of freedom of movement were required:

- **vergence**: the ability to bring the cameras to convergence such that there is zero disparity at the centre of the viewing volume, thus minimising the range of disparities in resulting stereograms;
- **elevation**: the ability to rotate the cameras in elevation so that the vergence point can be moved vertically within the working volume;
- **azimuth**: the ability to rotate the cameras in azimuth so that the vergence point can be moved horizontally within the working volume.

The range of movement for each axis was required to be at least sufficient to move the vergence point within the limits required by the working volume, although there would be obvious advantages if greater movement was available.

3.4.6. Speed of actuation

The speed of actuation was not a primary consideration for the ASP stereo vision system since the cameras were to be under the control of a human operator. Therefore the speed of actuation need only be commensurate with the relatively slow speed of human reactions as opposed to the potentially fast reactions of an automatic controller. The capability to move the vergence point to any point inside the limits of the working volume within the order of a few seconds was deemed to be sufficient, although it would of course be advantageous if quicker movement was possible.

3.5. Conclusions

The principal objective of the work described in this thesis was to develop the 3D computer vision components of the ASP system. An additional objective was to integrate these components into a complete stand-alone stereo 3D computer vision system so that the components could be tested and demonstrated in isolation from the full ASP system. The proposed approach for achieving these objectives can be summarised as follows:

1. *Development of a model of the proposed ASP sensor head.* In order to allow the interaction between the imaging and kinematic components of the stereo sensor

head to be analysed, it was proposed to develop the ASP sensor system model. In particular, this would allow the analysis experiments described in items 2 and 3 below to be performed.

2. *Analysis and selection of a basic configuration for the ASP sensor head.* In order to select a suitable basic configuration for the ASP sensor head, a series of simulation experiments using the ASP sensor system model was proposed. These experiments would analyse the relationship between the primary stereo vision system parameters (camera resolution, lens focal length, stereo baseline and object distance) and the basic performance of the system (working volume, disparity limit and resolution).

3. *Analysis of sensor head tolerances.* In order to characterise the effect of mechanical tolerances in the ASP sensor head a second series of novel simulation experiments using the ASP sensor system model was proposed. This would allow the level of surface measurement accuracy achievable using the proposed direct feedback approach to calibration to be determined. Hence it would be possible to design and fabricate the ASP sensor head so that it could provide the desired level of performance.

4. *Stereo matcher performance improvement.* A drastic improvement in the speed of execution of the MSSM stereo matcher was required. It was proposed that this would be achieved by analysing the MSSM stereo matching algorithm to identify possible performance improvements. The most promising sources of improvement would then be implemented.

5. *Stereo matcher performance characterisation.* The performance characteristics of the new version of the MSSM stereo matching algorithm would be unknown. Therefore, it was proposed to address this problem by carrying out a series of experiments to characterise its performance in general, and its accuracy in particular.

6. *Development of a static calibration scheme.* A method was required to provide an initial estimate of the positions and orientations of the cameras in the ASP sensor head when they are held in static positions. It was proposed to devise, implement and analyse the accuracy of a static calibration scheme that would be suitable for providing the required system calibration data.

7. *Development of an open loop dynamic calibration scheme.* The ASP system had to be capable of maintaining calibration after sensor movement. It was proposed to

achieve this initially by investigating and implementing an open-loop dynamic calibration scheme, based upon direct positional feedback from the camera actuators. The accuracy achieved using this scheme would be analysed.

8. *Development of a closed-loop dynamic calibration scheme.* It was proposed to refine the system calibration estimates derived from static calibration and open-loop dynamic calibration using a bundle adjustment algorithm. This would require an algorithm suitable for use with ASP stereo vision system to be devised, implemented and tested.

9. *System integration.* A final requirement of the work was to test and demonstrate the stereo 3D computer vision components developed for the ASP system. It was proposed to do this by integrating these components together with a 3D visualisation module into a complete stand-alone system.

Chapter 4
Sensor system model

4.1. Introduction

4.1.1. Objective

The objective of the work described in this chapter was to analyse the ASP sensor system to provide a mathematical model of the sensor head for the proposed ASP dynamic stereo vision system, and then to develop a software simulator. Although it was desirable that the developed software model simulate a generic dynamic stereo vision system, it had to support the specific types of analysis to be carried out during the design of the ASP sensor head.

The experiments planned for chapter 5 required the ASP sensor system model to analyse the relationships between the fundamental parameters of a verged stereo vision system (e.g. camera resolution, camera lens focal length, stereo baseline and vergence point position) and the basic performance of the system in terms of viewing volume, amount of disparity and planimetric and depth resolution. In chapter 6 it was planned that the model be used to investigate the relationship between tolerances in the components and construction of the sensor system and the 3D measurement accuracy attainable using open-loop dynamic calibration. The fundamental objective of this chapter was to develop the ASP sensor system model so that it would provide support for these two sets of planned experiments.

4.1.2. Background

The methods used to design even a fixed-viewpoint stereo vision system have gained practically no coverage in the literature. This would seem to imply that, if any design analysis has been carried out at all, only the simplest of procedures have been used and that these procedures have consequently been deemed unworthy of publication. Furthermore, the apparently rudimentary nature of the available design procedures has permitted only the simplest viewing geometries to be considered.

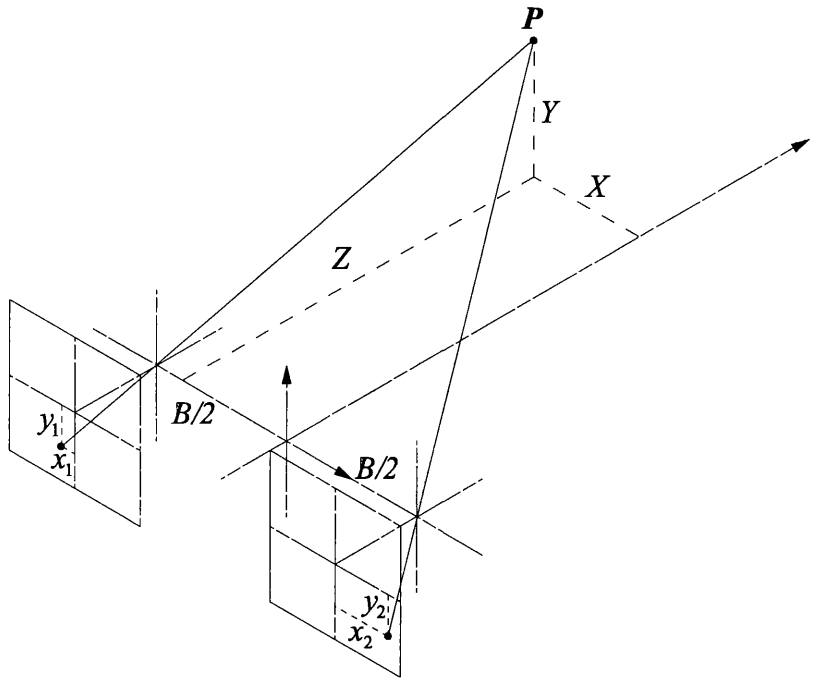


Figure 4.1. The "parallel" stereo viewing geometry.

The above argument is borne out by the popularity of the simplest possible viewing geometry for a stereo vision system: the "parallel" viewing geometry that is shown in figure 4.1. If the world co-ordinate frame is chosen to lie mid-way between the camera co-ordinate frames with the same orientation, as it has been in figure 4.1, then the relationship between the co-ordinates (X, Y, Z) of a world point, $\boldsymbol{P}$, and the Camera One and Camera Two image co-ordinates, (x_1, y_1) and (x_2, y_2) respectively, is:

$$x_1 = -c\frac{X + B/2}{Z} \tag{4.1}$$

$$x_2 = -c\frac{X - B/2}{Z} \tag{4.2}$$

$$y_1 = -c\frac{Y}{Z} \tag{4.3}$$

$$y_2 = -c\frac{Y}{Z} \tag{4.4}$$

where c is the camera principal distance and B is the baseline separation of the cameras. (It should be noted that equations 4.1 to 4.4 can be derived from the collinearity equations (equations 2.15 and 2.16) by substituting the values: $\omega = 0$, $\phi = 0$, $\kappa = 0$, $Y_0 = 0$, $Z_0 = 0$, and $X_0 = \pm B/2, -B/2$ for Camera One and $+B/2$ for Camera Two).

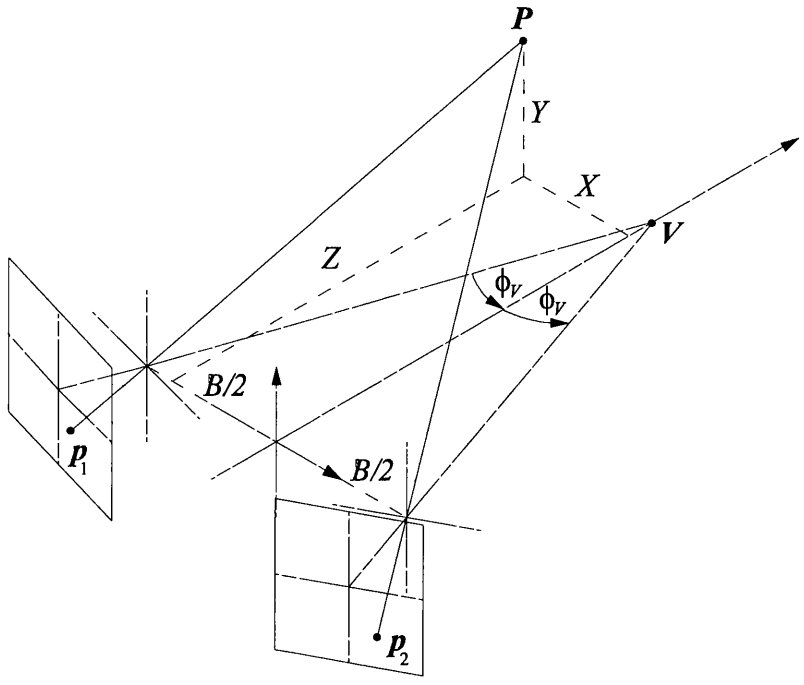


Figure 4.2. The "symmetrically verged" stereo viewing geometry.

Unfortunately, even a modest increase in the complexity of the stereo imaging geometry beyond the "parallel" viewing geometry greatly increases the complexity of the equations required to represent it. Consider the "symmetrically verged" stereo imaging geometry that is illustrated in figure 4.2. This imaging geometry is only slightly more complex than the "parallel" imaging geometry. The cameras are positioned in exactly the same way as for the "parallel" case and then rotated so that their optical axes converge on a point, ***V***, on the world *Z*-axis. To achieve this, Camera One and Camera Two are rotated by $+\phi_V$ and $-\phi_V$ respectively about their respective *Y*-axes. In this case, the collinearity equations (as described in section 2.2.2) become:

$$x_1 = -c\frac{(X + B/2)\cos\phi_V - Z\sin\phi_V}{(X + B/2)\sin\phi_V + Z\cos\phi_V} \tag{4.5}$$

$$x_2 = -c\frac{(X - B/2)\cos\phi_V + Z\sin\phi_V}{-(X - B/2)\sin\phi_V + Z\cos\phi_V} \tag{4.6}$$

$$y_1 = -c\frac{Y}{(X + B/2)\sin\phi_V + Z\cos\phi_V} \tag{4.7}$$

$$y_2 = -c\frac{Y}{-(X - B/2)\sin\phi_V + Z\cos\phi_V} \tag{4.8}$$

It is clear from equations 4.5 to 4.8 that the inclusion of even one extra parameter in the imaging geometry results in a considerable increase in the complexity of the equations required to represent it.

In order to meet the specified working volume requirements, the ASP sensor head will have to be able to assume a variety of asymmetric stereo imaging geometries with different azimuth and elevation (pan and tilt) rotations for each camera. The ASP sensor system model will not only have to accommodate this more complex imaging geometry, but it will also have to include additional parameters to simulate the dynamic nature of the cameras. For example, in order to support the experiments planned for Chapter 6, additional parameters will be required to simulate the rotation, position and tolerances of the individual rotational actuators. Not even the full formulation of the collinearity equations includes parameters of this kind. This is because the collinearity equations are intended to model only the final position and orientation of the cameras and not the mechanism by which this position and orientation is achieved.

The approach of directly deriving a few equations to model the imaging system that has apparently been adopted in the past for designing imaging systems may be adequate for the design of static stereo vision systems that employ simple imaging geometries such as "parallel" or even "symmetrically verged". However, it is clear from the above discussion, that this approach is not practical for modelling a vision system such as the proposed ASP stereo vision system that employs a more complex imaging geometry and animate sensors.

4.1.3. Approach

A new approach, inspired by techniques used in the fields of computer graphics and robotics, was devised for the development of the ASP sensor system model. In robotics, it is common practice to derive the position and orientation of a robot's end effector by applying a series of co-ordinate transformations for each of the robot's actuators in turn. It is often a requirement in computer graphics to synthesise the images produced by a virtual camera moving through a computer generated 3D environment. This is usually done by first computing the position and orientation of the camera at each viewpoint and then applying an "imaging transformation" to generate the required image from the 3D scene. The approach adopted for the ASP sensor system model was to apply the robotics technique to derive the position and orientation of the cameras first and then to apply the computer graphics imaging transformation.

The formulation of the mathematical model is therefore based upon the derivation of two basic types of co-ordinate transformation:

- the transformation of points between the 3D world co-ordinate system in which object points are defined and the 3D co-ordinate systems defined for each of the cameras;

- the perspective transformation between points in the 3D co-ordinate system associated with each camera and the camera's 2D image plane.

The mathematical model therefore naturally divides into two separate modelling components: a kinematic model simulating the former class of transformations and an imaging model simulating the latter class.

4.2. Kinematic model

4.2.1. Sensor head arrangement

Prior to developing the kinematic model of the ASP sensor head, it was first necessary to decide how the actuators and stereo cameras should be configured. Two basic arrangements were considered, each offering the four degrees of freedom specified in the system requirements. The first arrangement comprised an independent azimuth actuation for each of the cameras and the combined actuation of both cameras in azimuth and elevation. This arrangement is analogous to providing independent rotation of each eye in the horizontal plane and neck rotation in both horizontal and vertical planes. The second arrangement that was considered consisted of both independent azimuth and elevation actuation for each camera. This is analogous to providing an independent rotation of each eye in both horizontal and vertical planes, but with no provision for neck rotation.

Although the neck actuation offered by the first configuration would have provided the Active Stereo Probe with slightly greater functionality and flexibility, it was believed that it would also have hindered the integration of the projection system into the sensor head. A further disadvantage of the first configuration was considered to be the increased mechanical complexity of providing a two-axis actuation of a neck supporting both cameras, compared with the relative ease perceived in providing independent two-axis actuation for both cameras. For these reasons, the second configuration, which is illustrated in figure 4.3, was adopted.

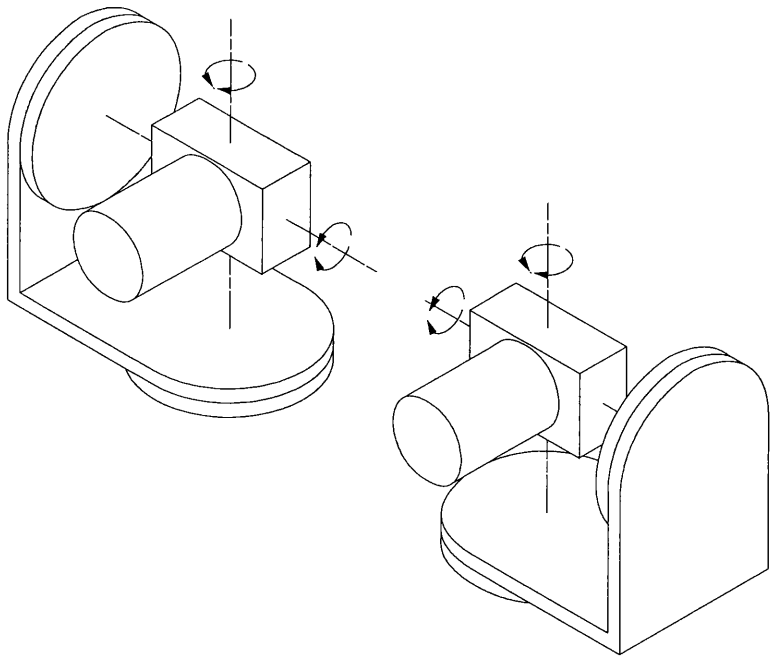


Figure 4.3. The configuration of cameras and actuators chosen for the ASP sensor head.

4.2.2. Homogeneous co-ordinate transformation

The primary mathematical tool used in the development of the kinematic model is the homogeneous transformation of co-ordinate systems. The homogeneous transformation of the co-ordinates of a point, ***P***, from co-ordinate system *A* to co-ordinate system *B* is carried out as follows [Sommerville 1951]:

$$\begin{bmatrix} {}^{B}X_P \\ {}^{B}Y_P \\ {}^{B}Z_P \\ 1 \end{bmatrix} = {}^{B}_{A}\mathbf{T} \begin{bmatrix} {}^{A}X_P \\ {}^{A}Y_P \\ {}^{A}Z_P \\ 1 \end{bmatrix} \tag{4.9}$$

where: $\left({}^{A}X_P, {}^{A}Y_P, {}^{A}Z_P\right)$ are the co-ordinates of ***P*** in co-ordinate system *A*, $\left({}^{B}X_P, {}^{B}Y_P, {}^{B}Z_P\right)$ are the co-ordinates of ***P*** in co-ordinate system *B* and ${}^{B}_{A}\mathbf{T}$ is the 4×4 homogeneous transformation matrix. Note that in the notation ${}^{B}_{A}\mathbf{T}$, a preceding subscript is used to denote the source co-ordinate frame and a preceding superscript is used to denote the destination co-ordinate frame. Note also that in the notation used to represent co-ordinates, e.g. $\left({}^{A}X_P, {}^{A}Y_P, {}^{A}Z_P\right)$, the preceding superscript denotes the co-ordinate system and the proceeding subscript identifies the point.

The transformations of co-ordinate systems in the kinematic model are built up from simple translations and rotations. Henceforth ***Tran***(X, Y, Z) will be used to represent a translation of (X, Y, Z), ***RotX***(Ω) will be used to represent a rotation of Ω about the X-axis, ***RotY***(Φ) will be used to represent a rotation of Φ about the Y-axis and ***RotZ***(K) will be used to represent a rotation of K about the Z-axis.

$$\boldsymbol{Tran}(X,Y,Z)=\begin{bmatrix}1 & 0 & 0 & X\\ 0 & 1 & 0 & Y\\ 0 & 0 & 1 & Z\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{4.10}$$

$$\boldsymbol{RotX}(\Omega)=\begin{bmatrix}1 & 0 & 0 & 0\\ 0 & \cos(\Omega) & \sin(\Omega) & 0\\ 0 & -\sin(\Omega) & \cos(\Omega) & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{4.11}$$

$$\boldsymbol{RotY}(\Phi)=\begin{bmatrix}\cos(\Phi) & 0 & -\sin(\Phi) & 0\\ 0 & 1 & 0 & 0\\ \sin(\Phi) & 0 & \cos(\Phi) & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{4.12}$$

$$\boldsymbol{RotZ}(\mathrm{K})=\begin{bmatrix}\cos(\mathrm{K}) & \sin(\mathrm{K}) & 0 & 0\\ -\sin(\mathrm{K}) & \cos(\mathrm{K}) & 0 & 0\\ 0 & 0 & 1 & 0\\ 0 & 0 & 0 & 1\end{bmatrix} \tag{4.13}$$

4.2.3. The viewing transformation

The purpose of the kinematic model is to derive the co-ordinate transformation between the world co-ordinate system and the co-ordinate systems of the stereo cameras. The kinematic model thus requires two homogeneous co-ordinate transformations, ${}^{C1}_{W}\mathbf{T}$ and ${}^{C2}_{W}\mathbf{T}$, for the transformation between the world co-ordinate system and each of the two camera co-ordinate systems. Since these two transformations are derived in an identical manner, the analysis has been simplified by dropping the camera identification suffix and considering only one such transformation, ${}^{C}_{W}\mathbf{T}$, which is equally applicable to both cameras.

The transformation ${}^{C}_{W}\mathbf{T}$ has been termed the *viewing transformation* because it transforms the world co-ordinates of a point, $\left({}^{W}X, {}^{W}Y, {}^{W}Z\right)$, into camera co-ordinates, $\left({}^{C}X, {}^{C}Y, {}^{C}Z\right)$. This viewing transformation must be applied to all world co-ordinate

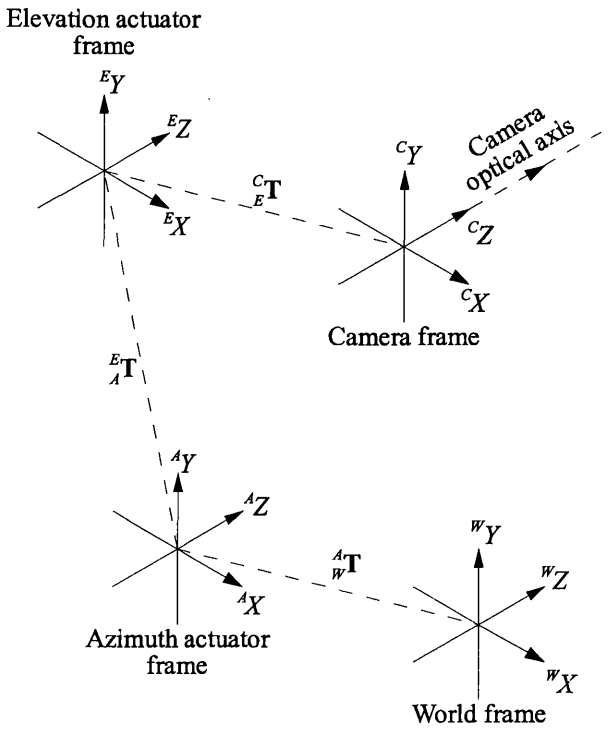


Figure 4.4. Co-ordinate frames for one actuated camera. Links are shown symbolically with dashed lines.

points before any of the *imaging transformations*, that are described later in this chapter, can be performed.

$$\begin{bmatrix} {}^{C}X \\ {}^{C}Y \\ {}^{C}Z \\ 1 \end{bmatrix} = {}^{C}_{W}\mathbf{T} \begin{bmatrix} {}^{W}X \\ {}^{W}Y \\ {}^{W}Z \\ 1 \end{bmatrix} \tag{4.14}$$

The Denavit-Hartenberg notation that is commonly used for modelling and controlling robotic systems has been adopted to facilitate the derivation of the viewing transformation for any given geometry of the sensor head. Two intermediate co-ordinate frames, associated with the azimuth and elevation actuators respectively, have been introduced between the world and camera frames. Links were then established between successive co-ordinate frames as shown in figure 4.4. The order in which these were introduced was designed to reflect the construction of the sensor head, i.e. first between the world frame and the azimuth actuator frame, then between the azimuth actuator frame and the elevation actuator frame and finally between the

elevation actuator frame and the camera frame. Each of these links is represented mathematically by the homogeneous transformation matrices ${}^{A}_{W}\mathbf{T}$, ${}^{E}_{A}\mathbf{T}$ and ${}^{C}_{E}\mathbf{T}$ respectively. This formulation allows the viewing transformation to be built up in a modular fashion by considering each link independently:

$${}^{C}_{W}\mathbf{T}={}^{C}_{E}\mathbf{T}\,{}^{E}_{A}\mathbf{T}\,{}^{A}_{W}\mathbf{T} \tag{4.15}$$

The composition of each of the link transformation matrices, ${}^{A}_{W}\mathbf{T}$, ${}^{E}_{A}\mathbf{T}$ and ${}^{C}_{E}\mathbf{T}$, was derived from a number of kinematic parameters of the elevation actuator, the azimuth actuator and camera respectively. These parameters are described in more detail in the following three sections.

4.2.4. The world-azimuth actuator link

The world-azimuth actuator link transformation matrix, ${}^{A}_{W}\mathbf{T}$, was derived using the following parameters (see figure 4.5, figure 4.6 and figure 4.7):

- The world co-ordinate location of the azimuth actuator origin $\left({}^{W}X_{A},{}^{W}Y_{A},{}^{W}Z_{A}\right)$.
- An angle of rotation, ϕ_{A}, about the azimuth actuator Y-axis and another angle of rotation, ω_{A}, about the azimuth actuator X-axis required to model tilt, or wobble, of the plane of the primary rotation of the azimuth actuator.
- The primary rotation performed by the azimuth actuator expressed as an angle of rotation, Φ_{A} about the azimuth actuator Y-axis.

The full formulation of the world-azimuth actuator link transformation is then:

$${}^{A}_{W}\mathbf{T} = \boldsymbol{RotY}\left(\phi_{A}\right)\boldsymbol{RotX}\left(-\omega_{A}\right)\boldsymbol{RotY}\left(-\Phi_{A}-\phi_{A}\right)\boldsymbol{Tran}\left(-{}^{W}X_{A},-{}^{W}Y_{A},-{}^{W}Z_{A}\right) \tag{4.16}$$

4.2.5. The azimuth actuator-elevation actuator link

The link between the azimuth actuator and elevation actuator, ${}^{E}_{A}\mathbf{T}$, is derived using a similar set of parameters to those used in the derivation of the world-azimuth actuator link:

- The azimuth actuator co-ordinate location of the elevation actuator origin $\left({}^{A}X_{E},{}^{A}Y_{E},{}^{A}Z_{E}\right)$.

Figure 4.7. Rotation of the azimuth actuator co-ordinate frame, Φ_A.

Figure 4.6. Off axis wobble of the azimuth actuator co-ordinate frame, ϕ_A and ω_A.

Figure 4.5. Translation between the world and azimuth actuator co-ordinate frames, (^WX_A, ^WY_A, ^WZ_A).

- The rotation performed by the elevation actuator expressed as an angle of rotation, Ω_E about the elevation actuator X-axis.

- An angle of rotation, ω_E, about the elevation actuator X-axis and another angle of rotation, κ_E, about the elevation actuator Z-axis required to model off-axis tilt or wobble of the elevation actuator.

The full formulation of the azimuth actuator-elevation actuator link transformation is:

$$ {}^{E}_{A}\mathbf{T} = \boldsymbol{RotX}(\omega_E)\boldsymbol{RotZ}(-\kappa_E)\boldsymbol{RotX}(-\Omega_E - \omega_E)\boldsymbol{Tran}(-{}^{A}X_E, -{}^{A}Y_E, -{}^{A}Z_E) \quad (4.17) $$

4.2.6. The elevation actuator-camera link

Unlike the world-azimuth actuator link and the azimuth actuator-elevation actuator link, there is no actuation component in the link between the elevation actuator and the camera. However a consistent approach to deriving the link transformation matrices was maintained by including the rotation of the camera about its optical axis and off-axis tilt of the camera in the model. The six parameters used to define the elevation actuator-camera link transformation were therefore:

- The elevation actuator co-ordinate location of the origin of the camera co-ordinate frame, $({}^{E}X_C, {}^{E}Y_C, {}^{E}Z_C)$.

- The rotation of the camera about its optical axis expressed as an angle of rotation, K_C about the camera Z-axis.

- An angle of rotation, κ_C, about the camera Z-axis and another angle of rotation, ϕ_C, about the camera Y-axis that model off-axis tilt or wobble of the camera.

The full formulation of the elevation actuator-camera link transformation matrix is therefore:

$$ {}^{C}_{E}\mathbf{T} = \boldsymbol{RotZ}(\kappa_C)\boldsymbol{RotY}(-\phi_C)\boldsymbol{RotZ}(-K_C - \kappa_C)\boldsymbol{Tran}(-{}^{E}X_C, -{}^{E}Y_C, -{}^{E}Z_C) \quad (4.18) $$

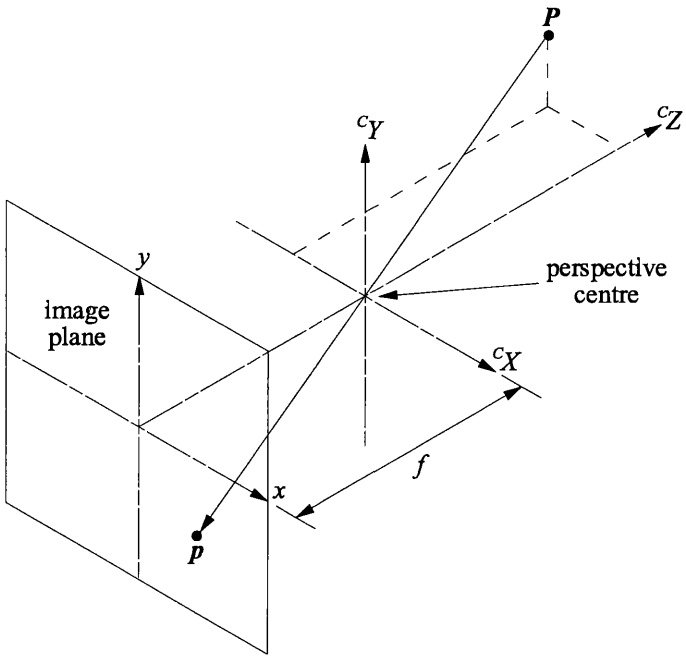


*Figure 4.8. Perspective projection of a world point, **P**, onto an image point, **p**.*

4.3. Imaging model

4.3.1. Image formation

The basic function of the imaging model is to model the image formation process, i.e. to determine the perspective projection of 3D world points onto 2D image points. The camera co-ordinate system is defined such that the ${}^{C}Z$-axis is the optical axis of the camera lens and the perspective centre of a camera is the origin of the camera co-ordinate system. Furthermore, the imaging plane of the camera is the plane ${}^{C}Z = -c$, where c is the principal distance of the camera lens assuming the thin lens approximation, and the image x- and y-axes are parallel to the camera co-ordinate ${}^{C}X$- and ${}^{C}Y$-axes. The camera co-ordinate and image co-ordinate systems were defined in this way so that the image point, $\boldsymbol{p}$, of an object point, $\boldsymbol{P}$, can be determined using the standard equations for thin lens perspective projection:

$$x_p = \frac{-c\ {}^{C}X_P}{{}^{C}Z_P} \tag{4.19}$$

$$y_p = \frac{-c\ {}^{C}Y_P}{{}^{C}Z_P} \tag{4.20}$$

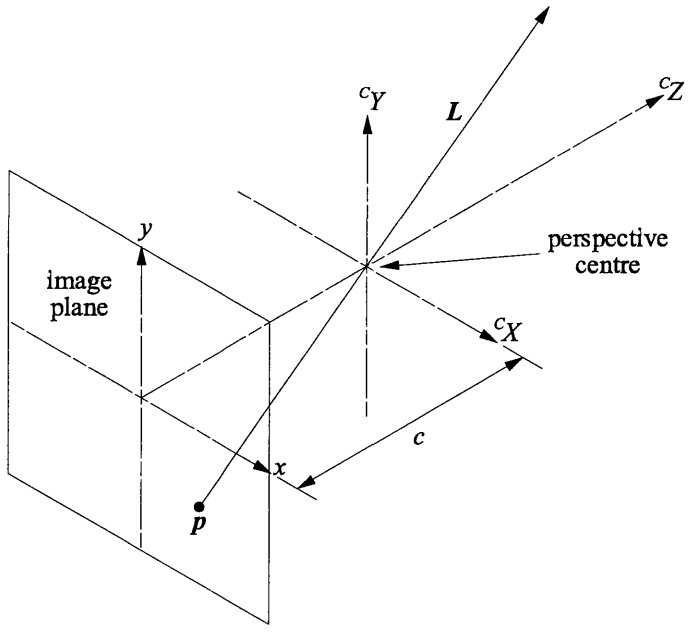


*Figure 4.9. Back projection of an image point, **p**, to a space line, **L**.*

Where (x_p, y_p) are the image co-ordinates of $\boldsymbol{p}$ and $({}^{C}X_P, {}^{C}Y_P, {}^{C}Z_P)$ are the camera co-ordinates of $\boldsymbol{P}$. Figure 4.8 illustrates the perspective projection of $\boldsymbol{P}$ onto $\boldsymbol{p}$. In the model, the co-ordinates of object points will generally be specified in the world co-ordinate system. It will be necessary, therefore, to perform the viewing transformation, ${}^{C}_{W}\mathbf{T}$, prior to performing perspective projection.

4.3.2. Back projection

Back projection, the inverse of perspective projection, is the process of projecting image points back into three dimensions. This is the first step in either the recovery of the three-dimensional world location of a point from a stereo pair of image points or the derivation of the epipolar line in one image from an image point in the other. These respective processes are described in sub-sections 4.3.3 and 4.3.4 that follow. Since back projection involves a transformation from two dimensions into three dimensions, the co-ordinate location of the resulting point cannot be determined in full. It can be seen from figure 4.9 that the result is instead a line in three-dimensional space that passes through the perspective centre of the camera and the location of the image point on the camera image plane.

From the previous definition of the image and camera co-ordinate systems, the camera co-ordinates of an image point $\boldsymbol{p} = (x_p, y_p)$, are $(x_p, y_p, -c)$ and the perspective centre is the origin of the camera co-ordinate system, (0,0,0). The space line $\boldsymbol{L}$

formed by the back projection of $\boldsymbol{p}$ can therefore be represented in camera co-ordinates as the 4×2 homogeneous matrix:

$$ {}^{C}\mathrm{L} = \begin{bmatrix} x_p & 0 \\ y_p & 0 \\ -c & 0 \\ 0 & 1 \end{bmatrix} \tag{4.21} $$

The back projected line $\boldsymbol{L}$ can be transformed from the camera co-ordinate system, ${}^{C}\boldsymbol{L}$, to the world co-ordinate system, ${}^{W}\boldsymbol{L}$, using the inverse of the viewing transformation:

$$ {}^{W}\boldsymbol{L} = \left({}^{C}_{W}\mathbf{T}\right)^{-1} {}^{C}\boldsymbol{L} \tag{4.22} $$

4.3.3. Stereo back projection

Figure 4.10 shows that a stereo-pair of image points $\boldsymbol{p}$ and $\boldsymbol{q}$ on the image plane of Camera One and Camera Two respectively will back project to two space lines $\boldsymbol{L}$ and $\boldsymbol{M}$. By applying equations 4.21 and 4.22 to the image co-ordinates of $\boldsymbol{p}$ and $\boldsymbol{q}$, equations for the world co-ordinates of the space lines ${}^{W}\boldsymbol{L}$: and ${}^{W}\boldsymbol{M}$ can be derived:

$$ {}^{W}\mathrm{L} = \begin{bmatrix} \alpha_L & A_L \\ \beta_L & B_L \\ \gamma_L & C_L \\ 0 & 1 \end{bmatrix} = \left({}^{C1}_{W}\mathbf{T}\right)^{-1} \begin{bmatrix} {}^{1}x_p & 0 \\ {}^{1}y_p & 0 \\ -c_1 & 0 \\ 0 & 1 \end{bmatrix} \tag{4.23} $$

$$ {}^{W}\mathrm{M} = \begin{bmatrix} \alpha_M & A_M \\ \beta_M & B_M \\ \gamma_M & C_M \\ 0 & 1 \end{bmatrix} = \left({}^{C2}_{W}\mathbf{T}\right)^{-1} \begin{bmatrix} {}^{2}x_q & 0 \\ {}^{2}y_q & 0 \\ -c_2 & 0 \\ 0 & 1 \end{bmatrix} \tag{4.24} $$

The world point, $\boldsymbol{R}$, from which the stereo-pair of image points, $\boldsymbol{p}$ and $\boldsymbol{q}$, were projected can be recovered by finding the intersection of the space lines $\boldsymbol{L}$ and $\boldsymbol{M}$. Rewriting equations 4.23 and 4.24 in parametric form yields:

$$ \begin{aligned} {}^{W}X_L &= \mu_L \alpha_L + A_L \\ {}^{W}Y_L &= \mu_L \beta_L + B_L \\ {}^{W}Z_L &= \mu_L \gamma_L + C_L \end{aligned} \tag{4.25} $$

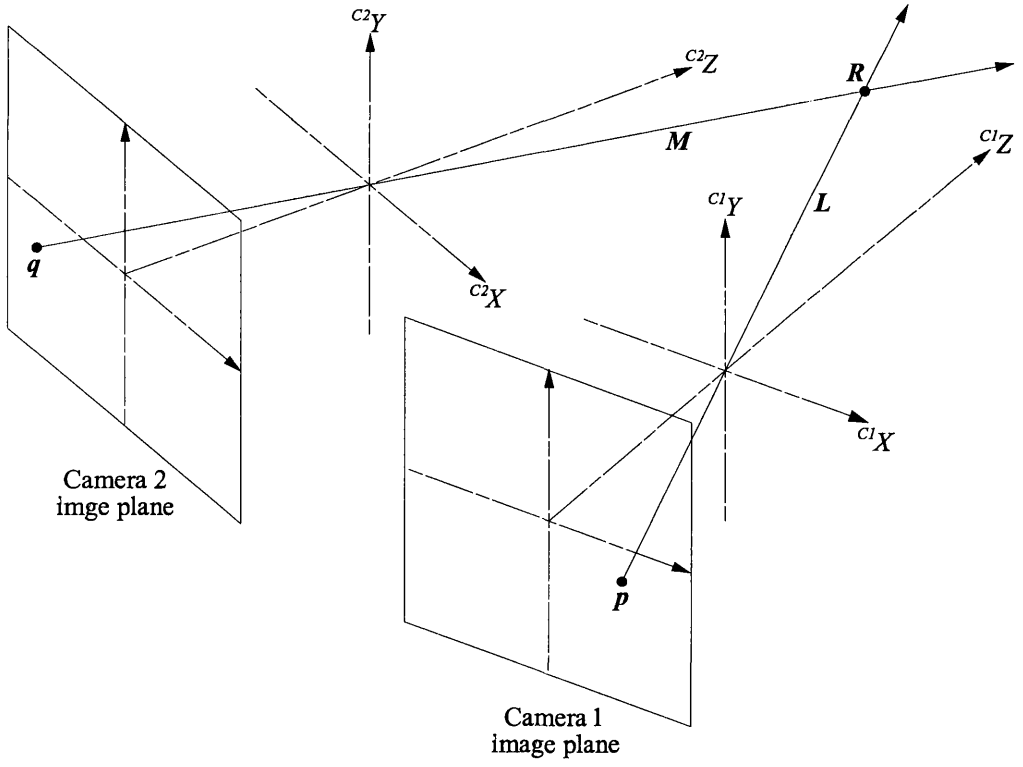


*Figure 4.10. A stereo pair of image points, **p** and **q**, back project as space lines, **L** and **M**, that intersect at a space point **R**.*

$$
\begin{aligned}
{}^{W}X_M &= \mu_M \alpha_M + A_M \\
{}^{W}Y_M &= \mu_M \beta_M + B_M \\
{}^{W}Z_M &= \mu_M \gamma_M + C_M
\end{aligned}
\tag{4.26}
$$

where $({}^{W}X_L, {}^{W}Y_L, {}^{W}Z_L)$ and $({}^{W}X_M, {}^{W}Y_M, {}^{W}Z_M)$ are the world co-ordinates of any point on the lines ***L*** and ***M*** respectively. The world co-ordinates of ***R*** can now be determined by equating the components of equations 4.25 and 4.26. This yields three equations in two unknowns; the unknowns being the parameters μ_L and μ_M. In the model, the complexity of solving this over-constrained set of equations was avoided by equating only the ${}^{W}X$ and ${}^{W}Z$ terms in equations 4.25 and 4.26. This is equivalent to finding the intersection of ***L*** and ***M*** when they are projected onto the ${}^{W}X$-${}^{W}Z$ plane.

The values of the parameters μ_L and μ_M are then substituted back into either equation 4.25 or equation 4.26 to find ${}^{W}X_R$ and ${}^{W}Z_R$, the world ${}^{W}X$- and ${}^{W}Z$-co-ordinates of ***R***. The mean of ${}^{W}Y_L$ and ${}^{W}Y_M$ produced from equations 4.25 and 4.26 is used as the value of ${}^{W}Y_R$, the ${}^{W}Y$-co-ordinate of ***R***.

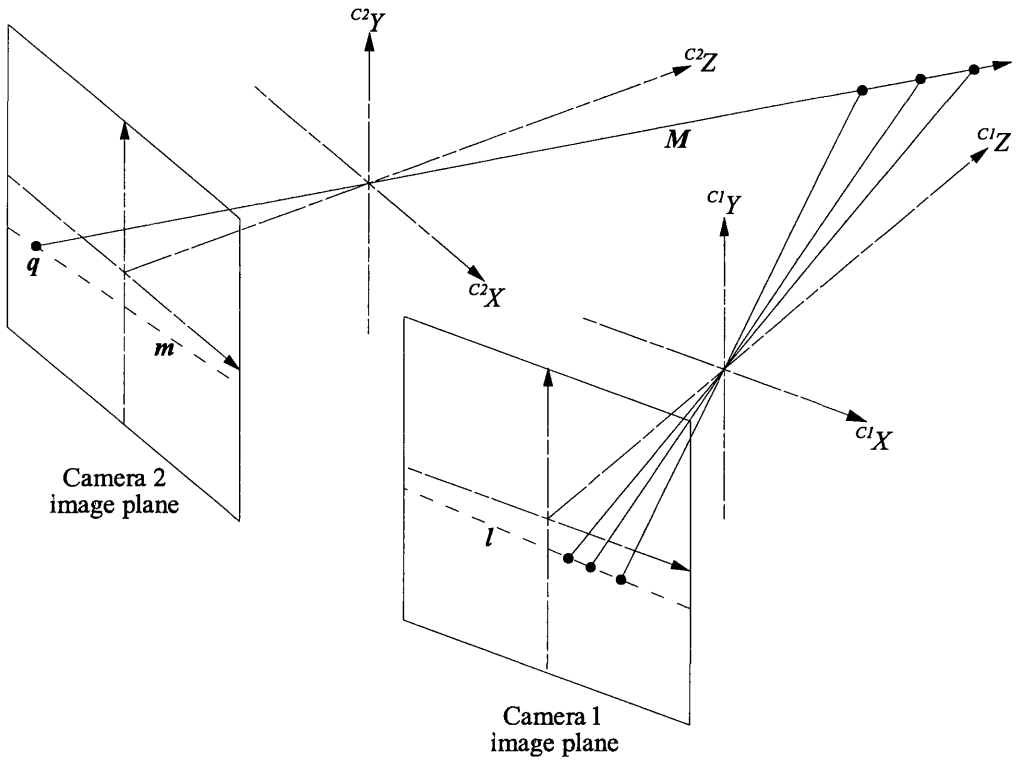


*Figure 4.11. Formation of an epipolar line, **l**, on the image plane of Camera One corresponding to an image point, **q**, on the image plane of Camera Two.*

4.3.4. Epipolar line derivation

An image point ***q*** on the image plane of Camera Two back-projects to a space line ***M*** as described earlier. If the space line ***M*** is then projected onto the image plane of Camera One, the result is a straight line ***l*** known as an *epipolar line*. The epipolar line ***l*** is the set of all points on the image plane of Camera One that can match the Camera Two image point ***q***. Figure 4.11 illustrates the derivation of an epipolar line.

Note that image point ***q*** also lies on an epipolar line ***m***. In fact, ***l*** is the set of all possible match points for any point on ***m***, and, conversely, ***m*** is the set of all possible match points for any point on ***l***. In photogrammetry, this relationship is known as the coplanarity constraint, since the epipolar lines ***l*** and ***m***, the perspective centres of the two cameras, and all possible world points corresponding to image points on ***l*** and ***m*** all lie on a plane.

The derivation of epipolar lines for a given stereo imaging configuration is important because their position and orientation will indicate the maximum horizontal and

vertical disparity than can be expected in captured stereo images. This is especially useful if it is possible to ascertain what portion of the epipolar line corresponds to a particular viewing volume. It is then possible to determine how much disparity can be expected from that combination of viewing volume and imaging configuration. Conversely, the derivation of epipolar lines will also allow the design of an imaging configuration that satisfies the disparity limits of the available stereo matcher while still managing to image the entire viewing volume.

In the ASP sensor system model, the epipolar line is derived by first back projecting image point $\boldsymbol{q}$ to the space line $\boldsymbol{M}$ as described earlier. The space line $\boldsymbol{M}$ is transformed from Camera Two co-ordinates to world co-ordinates and then from world co-ordinates to Camera One co-ordinates:

$$ {}^{C1}\mathrm{M} = \begin{bmatrix} \alpha & A \\ \beta & B \\ \gamma & C \\ 0 & 1 \end{bmatrix} = {}^{C1}_{W}\mathbf{T}\left({}^{C2}_{W}\mathbf{T}\right)^{-1} \begin{bmatrix} {}^{2}x_q & 0 \\ {}^{2}y_q & 0 \\ -c_2 & 0 \\ 0 & 1 \end{bmatrix} \tag{4.27} $$

By projecting any point on the Camera One co-ordinate line ${}^{C1}\boldsymbol{M}$ with ${}^{C1}Z_M > 0$ (i.e. excluding all points that lie behind Camera One) the homogeneous matrix for the epipolar line is derived as:

$$ {}^{1}\mathrm{l} = \begin{bmatrix} \dfrac{c_1(\alpha C - \gamma A)}{\gamma} & -\dfrac{c_1 A}{\gamma} \\ \dfrac{c_1(\beta C - \gamma B)}{\gamma} & -\dfrac{c_1 B}{\gamma} \\ 0 & 1 \end{bmatrix} \tag{4.28} $$

An exception to equation 4.28 exists if $\gamma = 0$, implying that $\boldsymbol{M}$ lies entirely within the ${}^{C1}Z = C$ plane. In this (unlikely) situation, the equation for the epipolar line becomes:

$$ {}^{1}\mathrm{l} = \begin{bmatrix} \dfrac{c_1\alpha}{C} & -\dfrac{c_1 A}{C} \\ \dfrac{c_1\beta}{C} & -\dfrac{c_1 B}{C} \\ 0 & 1 \end{bmatrix} \tag{4.29} $$

Furthermore, if $\gamma = 0$ and $C \leq 0$, then $\boldsymbol{M}$ lies within a plane ${}^{C1}Z = C$ which is behind the principal point of Camera One. There is therefore no valid epipolar line in this situation.

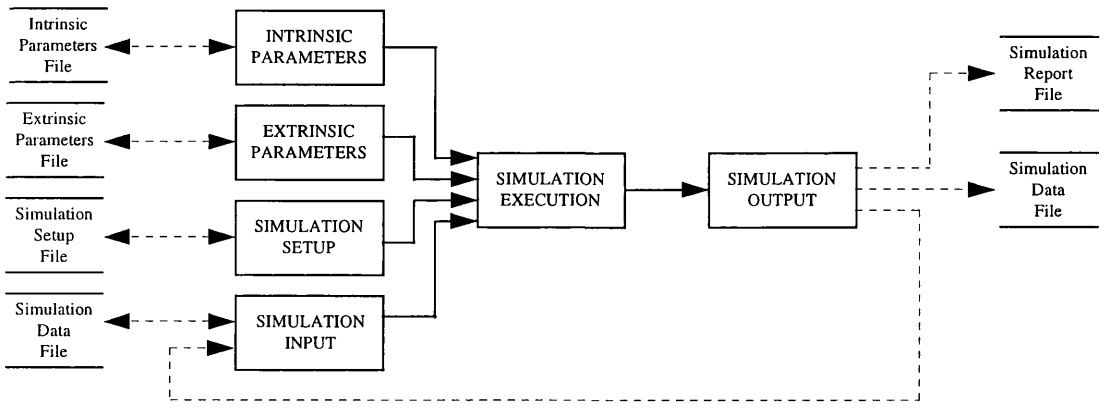


Figure 4.12. Schematic of the software implementation of the ASP sensor system model showing the six main modules boxed.

```
                HEAD MODEL MAIN MENU

        1. Camera Intrinsic Parameters Menu
        2. Camera Extrinsic Parameters Menu
        3. Simulation Setup Menu
        4. Simulation Input Menu
        5. Simulation Execution Menu
        6. Redraw The Screen

        7. QUIT

        Selection:
```

Figure 4.13. ASP sensor system model main menu screen.

4.4. Software implementation

The complete ASP sensor system model that combines the kinematic and imaging models described earlier in this chapter was implemented in the C programming language. The software implementation consists of the six main modules that are illustrated in figure 4.12. Each of these modules is described briefly below.

A user interface for the model was developed using the standard UNIX `curses` library so that it would be compatible with the widest possible variety of computers. The main menu from the user interface is shown in figure 4.13.

4.4.1. Camera intrinsic parameters module

Figure 4.14 shows the display screen for the intrinsic parameters module of the ASP sensor system model. This module maintains a list of the intrinsic parameters for each

```
                    INTRINSIC CAMERA PARAMETERS

A. Camera 1: Sony XC77CE Camera with 50mm lens
A. Camera 2: Sony XC77CE Camera with 50mm lens

                              Camera 1          Camera 2

B. Focal Length (mm):         50.000000         50.000000
C. Pixel Height (um):         10.890000         10.890000
D. Pixel Width (um):          10.780000         10.780000
E. Number of Rows:            576               576
F. Number of Columns:         768               768
G. Row Offset (pixels):       -287.500000       -287.500000
H. Column Offset (pixels):    -383.500000       -383.500000
```

Figure 4.14. ASP sensor system model camera intrinsic parameter display screen.

camera. These lists are passed to the simulation execution module when a simulation is to be run. The following basic functions can be applied to intrinsic parameters lists:

- *Edit:* allows any of the intrinsic parameters (comment string, focal length, pixel height, pixel width, number of rows, number of columns, row offset or column offset) for either camera to be given a new value.
- *Save:* the intrinsic parameter list for either camera can be saved to a file.
- *Load*: the intrinsic parameter list for either camera can be loaded into the model from a previously saved or pre-generated intrinsic parameters file.

4.4.2. Camera extrinsic parameters module

The display screen for the extrinsic parameters module of the ASP sensor system model is shown in figure 4.14. This module maintains a list of the values for the eighteen external parameters for each camera that are described in sub-sections 0, 4.2.5 and 4.2.6; i.e. three components of translation, a rotation angle, and an off-axis wobble angle and orientation for the azimuth actuator, the elevation actuator and the camera. In addition, the co-ordinates of a fixation point can be specified separately for each camera as described below (a verged stereo geometry can be obtained by specifying the same fixation point for both cameras). A comment string can also be added to the extrinsic parameter list for each camera. The following basic facilities are provided for manipulating the extrinsic parameter list:

- *Edit:* the value of any of the eighteen extrinsic parameters or the comment string for each camera can be given a new value.

```
                    EXTRINSIC CAMERA PARAMETERS

A. Camera 1: Default Extrinsic Parameters
A. Camera 2: Default Extrinsic Parameters

                                Camera 1            Camera 2

B. Azimuth X (mm):              -150.000000         +150.000000
C. Azimuth Y (mm):              +75.000000          +75.000000
D. Azimuth Z (mm):              +0.000000           +0.000000
E. Azimuth Rotation (Y):        -1.718367           +1.718367
F. Azimuth Wobble (Y):          +0.000000           +0.000000
G. Azimuth Wobble (X):          +0.000000           +0.000000

H. Elevation X (mm):            -30.000000          +30.000000
I. Elevation Y (mm):            -75.000000          -75.000000
J. Elevation Z (mm):            +0.000000           +0.000000
K. Elevation Rotation (X):      +0.000000           +0.000000
L. Elevation Wobble (X):        +0.000000           +0.000000
M. Elevation Wobble (Z):        +0.000000           +0.000000

N. Camera X (mm):               +30.000000          -30.000000
O. Camera Y (mm):               +0.000000           +0.000000
P. Camera Z (mm):               +50.000000          +50.000000
Q. Camera Rotation (Z):         +0.000000           +0.000000
R. Camera Wobble (Z):           +0.000000           +0.000000
S. Camera Wobble (Y):           +0.000000           +0.000000

Camera 1 Fixation Point:        (0.000000, 0.000000, 5.000000)
Camera 2 Fixation Point:        (0.000000, 0.000000, 5.000000)
```

Figure 4.15. ASP sensor system model camera extrinsic parameter display screen.

- *Fixate:* a new fixation point can be specified for either camera causing that camera's azimuth and elevation rotation parameters to change so that the camera is fixated on the specified world point.
- *Save:* the camera extrinsic parameter list for either camera can be saved to a file.
- *Load*: the extrinsic parameter list for either camera can be loaded into the model from a previously saved or pre-generated camera extrinsic parameters file.

4.4.3. Simulation setup module

A useful feature of the software implementation of the ASP sensor system is the ability to perform batch execution of a simulation using a range of values for selected intrinsic or extrinsic parameters. This feature of the model is controlled by the simulation setup module which maintains a list of the step size and number of steps to be applied to each intrinsic and extrinsic parameter during a batch simulation run. An example of a batch simulation run might be to repeat one of the simulations described

in sub-section 4.4.5 for a range of camera baseline separations from 100mm to 300mm in 10mm steps.

If a range of values are specified for two or more of the parameters, then the default operation is to perform a separate simulation cycle for all of the possible combinations. However it is also possible to "link" parameters together so that both are incremented at the same time during the simulation cycle. In the example above, the effect of repeating a simulation using baseline separations ranging from 100mm to 300mm in 10mm steps could be achieved by specifying a range of -50mm to -150mm in -5mm steps for the Camera One Azimuth X translation linked to a range of 50mm to 150mm in 5mm steps for the Camera Two Azimuth X translation.

The following functions are provided by the simulation setup module:

- *Add parameter:* adds a parameter to the list of parameters to be varied during the simulation. The step size, the number of steps and whether the parameter is linked to another must be specified.
- *Remove parameter:* removes a parameter from the list of parameters to be varied during the simulation.
- *Edit parameter:* changes the step size, number of steps and linkage of a previously specified parameter.
- *Save:* allows the simulation setup list to be saved to a file.
- *Load:* allows a previously saved simulation setup file to be loaded into the model.

4.4.4. Simulation input module

The simulation input module maintains a list of the current data that will be processed during the simulation. This simulation data can be in one of five possible formats:

- *3D points:* three-dimensional points, (X, Y, Z), in the world co-ordinate frame.
- *2D points:* two-dimensional image points, (x, y), in the co-ordinate frame of either of the cameras' image planes. Image point co-ordinates can be specified in either pixels (*column*, *row*) or in micrometres.
- *Stereo points:* stereo pairs of points, $((x_1, y_1), (x_2, y_2))$, in the co-ordinate frames of Camera One and Camera Two respectively. Stereo points can be specified in pixels or micrometres.

```
                              SIMULATION SETUP

     Baseline separation from 100mm to 300mm in 10mm steps

.....................................................................
     Param 1: Camera 1 Azimuth X Position [Not Linked]
         Current value =  -50.000000 mm
         Step size =       +5.000000 mm
         Positive steps = 0, Negative steps = 20

     Param 2: Camera 2 Azimuth X Position [Linked to 1]
         Current value =  +50.000000 mm
         Step size =       +5.000000 mm
         Positive steps = 20, Negative steps = 0
```

Figure 4.16. ASP sensor system model simulation setup display screen.

- *3D lines:* three-dimensional world co-ordinate lines stored in parametric form.
- 2D lines: two-dimensional lines in the co-ordinate frame of either of the cameras' image planes stored in parametric form.

The simulation input module allows the following functions to be performed on the simulation input data list:

- *Add data item:* allows a new point or line to be added to the list of simulation input data.
- *Remove data item:* removes a point or line from the current list of simulation input data.
- *Edit data item:* allows the co-ordinates or parameters of a point or line in the list of simulation input data to be changed.
- *Save:* allows the current list of simulation input data to be saved to a file.
- *Load:* allows a previously saved or pre-generated simulation data file to be loaded into the model.

4.4.5. Simulation execution module

The simulation execution module controls the simulation cycle. Five basic types of simulation can be chosen by the user:

- *Image formation:* projects every 3D point in the simulation input data onto a 2D image point on the image plane of the selected camera.

- *Back projection:* projects every 2D point in the simulation input data from the image plane of the selected camera to a 3D world co-ordinate line.

- *Back projection onto plane:* finds the 3D world point of intersection between the back projected line from every 2D point in the simulation input and a specified three-dimensional world co-ordinate plane.

- *Stereo back projection:* back projects every stereo pair of 2D image points from the image planes of the cameras to a 3D world co-ordinate point.

- *Epipolar derivation:* finds the 2D epipolar line on the image plane of a specified camera for every 2D point in the simulation input data on the image of the other camera.

The simulation cycle is repeated once for each step specified in the simulation setup list with the resulting data being appended to the current simulation output data list. The initial value for each intrinsic or extrinsic parameter is taken from the current intrinsic or extrinsic parameter lists respectively.

In addition to the five basic types of simulation provided in the ASP sensor system model, a number of composite simulations that perform two or more of the basic types of simulations were also implemented. For example, one composite simulation that was used extensively during the experiments reported in Chapters 5 and 6 first back projected 2D image points from one camera onto a world plane and then projected the resulting world points onto the image plane of the other camera.

4.4.6. Simulation output module

The simulation output module controls the processing of the output data from the simulation execution module. Three basic functions can be applied to the simulation output data list:

- *Save to data file:* saves the simulation output data list to a simulation data file that can be loaded into the simulation input module at a later date.

- *Save to report file:* generates a report file for the executed simulation. This details: the initial intrinsic and extrinsic parameters for each camera, the simulation setup, the simulation input data and the simulation output data formatted as a table.

- *Copy to input:* copies the simulation output data to the simulation input data list overwriting the existing simulation input data. The simulation output data can then be used as the input to a new simulation run.

4.5. Conclusions

The primary consideration in developing the ASP sensor system model was that it should provide support for the simulation experiments planned for designing the ASP sensor head. The ability of the developed model to meet this objective will be seen in Chapters 5 and 6 following.

The modular approach used within the ASP sensor system model has allowed the construction of a much more detailed model of a stereo vision system than any previous approach. The high level of detail provided by the model will be shown in the following chapters to allow a much more thorough design process to be carried out than for any previously reported dynamic stereo vision system. This is particularly true for the model's ability to simulate the effects of the tolerances in each individual actuator which will prove to be necessary for the experiments that are described in Chapter 6 and that are fundamental to the whole approach to quantitative dynamic stereo vision. These effects could simply not have been analysed and the proposed approach to quantitative dynamic stereo vision would not have been viable without the development of the ASP sensor system model.

A secondary objective of the work described in this chapter was to develop a stereo vision system model that would be as widely applicable as possible. The modular design of the ASP sensor system model has meant that, to a large degree, this objective has also been achieved successfully. The model has proved to be a useful tool in analysing aspects of the ASP stereo vision system beyond the system's design that was the primary consideration in developing the model. One example of this kind of analysis is the use of the model to help in the design of the calibration grid that is described in Chapter 10. More recently, the flexibility provided by the ASP stereo vision system has also allowed it to be used for the analysis of completely new stereo vision systems as described in Chapter 12.

In retrospect, the only major improvements that could be made to the ASP sensor system model would be in its implementation. Whenever a composite simulation comprising a chain of several of the basic simulation types was required, it was often necessary to add this into the C source code of the model and to re-compile it each time. This was found to be quite inconvenient and time consuming as well as restricting some of the inherent flexibility of the model. It was also found that the

menu driven user interface to the model made it awkward to specify batches of simulations, particularly if different simulation input data was required during each run. This problem was circumvented in practice by writing UNIX shell scripts to run the model several times and to provide on the standard input the equivalent key presses required for each run. However, the process of determining and specifying the required key presses was found to be rather time consuming and painstaking.

Given the software that was available when the model was first developed, it would have taken considerable effort at that time to address the relatively minor shortcomings described in the previous paragraph and still maintain the functionality and flexibility that was provided in the implementation. If the model was to be re-implemented now, the proper use of a modern scripting language such as Tcl [Ousterhaut, 1994] would not only address the above issues but provide even greater flexibility. If the basic functions of the model were provided as Tcl commands, it would be relatively simple to write scripts to perform batch processing and composite simulations. With careful design, it would also be possible to provide arbitrary numbers of cameras and different actuator arrangements. In addition an improved X-Windows user interface could easily be provided by using the Tk library of widgets that is integrated into Tcl.

Chapter 5
Sensor system configuration

5.1. Introduction

5.1.1. Objective

The objective of the experiments described in this chapter was to establish a basic configuration for the ASP sensor head that would be suitable for meeting the performance requirements specified in Chapter 3. The basic parameters that influence the performance of a stereo vision system are: the imaging characteristics of the stereo cameras, the principal distance of the camera lenses, the camera separation or baseline and the mean object distance. The performance factors that would be most directly influenced by the choice of these parameters would be the achievable viewing volume and the horizontal and vertical disparity limits.

The objective of the experiments described in this chapter was therefore to find a combination of principal distance, camera baseline and mean object distance that would provide the completed ASP stereo vision system with the required viewing volume and disparity limits. However, since the imaging characteristics of the cameras are so influential to the overall performance of a stereo vision system, the selection of the cameras for the ASP system was an essential preliminary stage to the experiments.

5.1.2. Stereo imaging geometry

Two basic stereo imaging configurations were considered for the ASP stereo vision system: a parallel stereo imaging geometry (as illustrated in figure 4.1 of chapter 4) or a convergent stereo imaging geometry (as illustrated in figure 4.2 of chapter 4). The preliminary experiments described in chapter 2 established that a convergent stereo imaging geometry offers two major advantages over a parallel one (overleaf):

1. It ensures that the mean of the disparity values obtained for all possible world points within the viewing volume is zero. This makes best use of the dynamic range of disparity provided by the stereo matcher, as described in Chapter 8. By contrast, stereograms captured using a parallel imaging geometry contain a fixed disparity offset. With many stereo matchers, including the MSSM stereo matcher, this must be removed before the stereograms can be matched.

2. It maximises the volume of overlap of the fields of view of the two cameras and hence provides the maximum viewing volume for a given pair of cameras and a given camera baseline. This is not the case for a parallel imaging geometry, which often provides significantly less overlap (over even no overlap in extreme cases!) between the cameras' fields of view.

However, a parallel stereo imaging geometry does offer one advantage over a convergent one: The epipolar lines in a parallel stereo imaging geometry are parallel lines which can be made to align with the scan lines of the sensing device. In this case, the disparity between corresponding stereo pairs of image points is purely horizontal. Therefore only a one dimensional search along a scan-line is required to find the stereo correspondence for each image point. A disadvantage of the convergent stereo imaging geometry is that it results in slanted epipolar lines (as illustrated in figure 4.11 of chapter 4), and thus introduces a vertical component of disparity. Consequently, a two dimensional search is required to find the correspondences in stereo images captured with convergent cameras. In the version of the MSSM stereo matching algorithm that was available at the start of this work, the number of calculations required to perform a 2D search was approximately the square of the number of calculations required to perform a 1D search. This clearly represents a considerable processing overhead. (This is discussed in more detail in chapter 8).

Nevertheless, the advantages gained from a convergent stereo imaging geometry were deemed to outweigh the disadvantage of the vertical component of disparity it introduces. It was therefore decided to employ a convergent stereo imaging geometry in the ASP sensor head, but to attempt to confine the amounts of vertical disparity to a level where a two dimensional search was not required (± 0.5 pixels was deemed to be an acceptable upper limit). Note that, unless stated otherwise, a convergent stereo imaging geometry should be assumed in the discussion presented in the remainder of this chapter.

5.1.3. Influencing factors

The objective of the work described in this chapter was to establish a basic configuration for the ASP stereo vision system that would meet the specified viewing volume and disparity limit requirements. An initial stage of this process was to select the cameras that would used. Assuming that the cameras had been selected, and hence that their imaging characteristics were fixed, the remaining parameters that would influence the ability of the stereo vision system to meet these requirements were:

- the principal distance of the lenses,
- the camera separation (or stereo baseline),
- the mean object distance (essentially the position of the vergence point).

The influence of these parameters on the amount of horizontal disparity that will result can be seen from the basic equation for a parallel stereo vision system:

$$\delta = \frac{cB}{Z} \tag{5.1}$$

This shows that (horizontal) disparity, δ, is directly proportional to principal distance, c, and baseline separation, B, and inversely proportional to distance, Z. Although this equation is not applicable exactly to the convergent stereo imaging geometry (as illustrated in figure 5.1), the amount of horizontal disparity will still vary approximately in proportion to lens principal distance and camera baseline separation and approximately in inverse proportion to distance.

As discussed in section 5.1.2 above, the vertical disparity introduced by a convergent stereo imaging geometry increases with the angle of convergence. If the angle of convergence, α, is defined as the angle between the optical axes of the cameras, then it is given by the equation:

$$\alpha = 2\tan^{-1}\left(\frac{B}{2Z}\right) \tag{5.2}$$

Therefore, the amount of vertical disparity will increase as the baseline is increased and decrease as the distance is increased.

Figure 5.1 shows that the width of the viewing volume is constrained by the overlap of the fields of view of the two cameras. The height of the viewing volume is also constrained by the fields of view of the cameras. Assuming that a particular camera

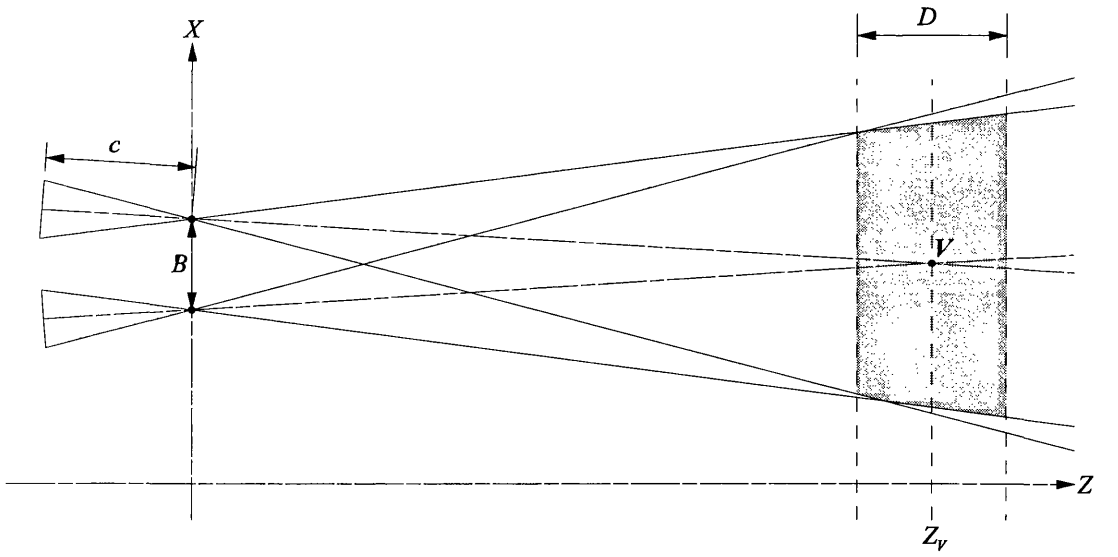


Figure 5.1. Plan view of a verged stereo vision system showing the viewing volume (shaded).

has already been selected, its angular field of view is controlled solely by the principal distance of its lens. Therefore, principal distance, c, is one factor that governs the viewing volume width and height.

Figure 5.1 also shows that the distance to the centre of the viewing volume is another factor governing its width and height. This is clearly the case since a camera's field of view expands with distance. Note that it is assumed that the vergence point, V, lies at the centre of the viewing volume and therefore that the distance to the vergence point is also the mean object distance, i.e. $Z_V = Z_{mean}$. It is also worthy of note that the dimensions of the viewing volume are not significantly influenced by the baseline separation of the cameras unless the angle of convergence is very large.

The depth of the viewing volume (D in figure 5.1) is not directly constrained by lens principal distance, mean object distance or baseline separation, unless the angle of convergence is very large. However it does influence strongly the required disparity limits of a stereo vision system. This is because, in a convergent stereo vision system, both horizontal and vertical disparity increase with distance from the vergence point. Therefore world points located at the front or back of a deep viewing volume will generate large (negative or positive respectively) disparities and hence will require a large disparity limit if they are to be successfully matched.

In summary, the viewing volume width and height of a convergent stereo system are influenced strongly by the lens principal distance and mean object distance and much

less by the camera baseline separation. By contrast, the required disparity limits are governed by all three of these factors and, in addition, by the required depth of the viewing volume.

5.1.4. Analytical approach

The huge range of potential configurations for the ASP sensor head and the complexity of determining whether a given configuration met the requirements made an exhaustive search for an "ideal" solution impractical. It was therefore decided to adopt a configuration likely to be close to a suitable solution as the starting point for the analysis. The parameters of this initial configuration would then be adjusted as necessary, in an iterative manner, until the specified performance requirements were met. Details of the chosen initial configuration are given in section 5.1.5.

The complexity of the analysis task was reduced further by taking the mean object distance directly from the initial trial configuration without modification and keeping it constant throughout the subsequent analysis. This would be possible since the lens principal distance and baseline separation alone would be capable of providing adequate control over the viewing volume and disparity limits. The centre of the working volume could then be defined as the world point situated at the mean object distance viewed directly in front of the cameras. For the subsequent analysis, the range of possible vergence points could then be constrained to lie within the 1.0m $\times$ 1.0m $\times$ 1.0m cube centred at this point. Thus, that part of the working volume specification would be met.

Having chosen the mean object distance and hence specified the bounding volume for vergence points, the next stage of the analysis was to determine the lens principal distance required to meet the specification for viewing volume width and height. The adopted approach was to use the ASP sensor model to analyse the dimensions of the viewing volume made possible for a range of lens principal distances about the initial value. The initial value for baseline separation would be used throughout this analysis and the cameras would be verged to the point at the centre of the working volume.

After selecting the most suitable principal distance suggested by this analysis, the viewing volume resulting at the extremes of the working volume would be tested using the model. If the results were found to be unacceptable, the principal distance could be adjusted appropriately. The process would then be repeated until a value of principal distance was found that would cause the viewing volume specification to be met satisfactorily.

The only parameter value remaining to be determined would thus be the camera baseline separation. A similar method was adopted to that used for determining the principal distance, except that the disparity limits resulting from a range of baselines about the initial value would be analysed. The principal distance would be set to the value just determined and the cameras would again be verged to the central vergence point. The disparity limits resulting from the most suitable baseline could then be tested at the extreme vergence points. Iteration of the process to find the most suitable baseline could then be carried out as necessary. If required, the whole process could also be iterated with the newly determined values for principal distance and baseline separation forming a new initial trial configuration.

5.1.5. Initial configuration

A major obstacle to the above approach is clearly the requirement to find a suitable initial configuration. Fortunately, just such a configuration was available: that of the early translating single camera stereo rig that was briefly described in Chapter 1. The experienced gained by the author in using this rather primitive system [Urquhart, 1989; Urquhart, 1990] had shown that it could meet quite closely the working volume and disparity limit requirements of the ASP system. The basic parameters of this system were:

- lens principal distance, $c = 50\text{mm}$;
- camera baseline, $B = 300\text{mm}$;
- mean object distance, $Z_{mean} = 5.0\text{m}$.

Unfortunately there were two notable differences between this stereo rig and the planned ASP sensor head. The single translating camera rig necessitated the use of a parallel imaging geometry as opposed to the verged geometry required for ASP stereo vision system. The standard Cohu video camera that was used in the single camera rig would not be suitable for use in the ASP system. Therefore, the cameras that would be chosen for use in the ASP system were likely to have different imaging characteristics to those of the Cohu camera. Nevertheless, these differences were considered to be slight enough that the combination of parameters listed above would make a suitable initial configuration for the following analysis.

Figure 5.2. A Sony XC77RR-CE remote camera head unit with miniature 50mm lens.

5.2. Choice of cameras

Sony XC77RR-CE remote head cameras, as illustrated in figure 5.2, were chosen for use in the Active Stereo Probe. These cameras were chosen because they offered the desired combination of full CCIR resolution, 576 × 768 pixels, and small size, 44mm × 31mm × 23mm, and weight, 65g (dimensions and weight for camera without lens or cable). It will be seen from the discussion later in this chapter that a square format camera would have been better able to meet the viewing volume requirements. However only analogue television cameras were available at the time when the cameras were being chosen and the best of these conformed to either the CCIR or NTSC standard image sizes (576 × 768 pixels or 480 × 640 pixels respectively).

The characteristics of the Sony XC77RR-CE camera that are important to the experiments in the rest of this chapter are given in table 5.1.

	Horizontal	Vertical
Pixel size (μm)	10.78	10.89
Sensor size (pixels)	768	576
Sensor size (mm)	8.8	6.6

Table 5.1. Notable characteristics of the Sony XC77RR-CE camera.

An additional factor favouring the Sony XC77RR-CE camera was the electronic shutter facility that it provided, and would be required for its use in conjunction with the flash based projection system. The Sony XC77RR-CE was chosen in preference to similar colour cameras because the significantly higher resolution provided by this monochrome camera was deemed to be more important than colour video output. (Colour images could be captured using monochrome cameras if three separate images are captured while the scene is illuminated sequentially with red, green and blue light).

5.3. Choice of mean object distance

The choice of 5.0m for the mean object distance may appear initially to be somewhat arbitrary. In fact, this is not the case, since there are several arguments to justify the choice of this value. These arguments essentially reduce to the selection between either a short object distance and a short principal distance or a long object distance and a long principal distance. Either combination would meet the specified viewing volume width and height requirements at the central vergence point. This can be seen by comparing figure 5.1, which illustrates a far vergence point, i.e. large Z_V, and a long principal distance, c, with figure 5.3, which illustrates a closer vergence point and a shorter principal distance.

The most important factor affecting the choice of mean object distance was the effect that this would have on the MSSM stereo matching algorithm. Section 5.1.2 explained that, although a convergent imaging geometry was required, it was desirable that this should also be as close to parallel as possible in order to reduce or even remove the requirement for the matcher to perform a slow 2D search. Therefore the favoured option was to employ as long an object distance as possible since this would result in as small an angle of convergence as possible and hence a stereo imaging geometry as close to parallel as possible.

It was also a requirement of the ASP stereo vision system that an adequate viewing volume be obtained for any vergence point within the 1.0m × 1.0m × 1.0m cube

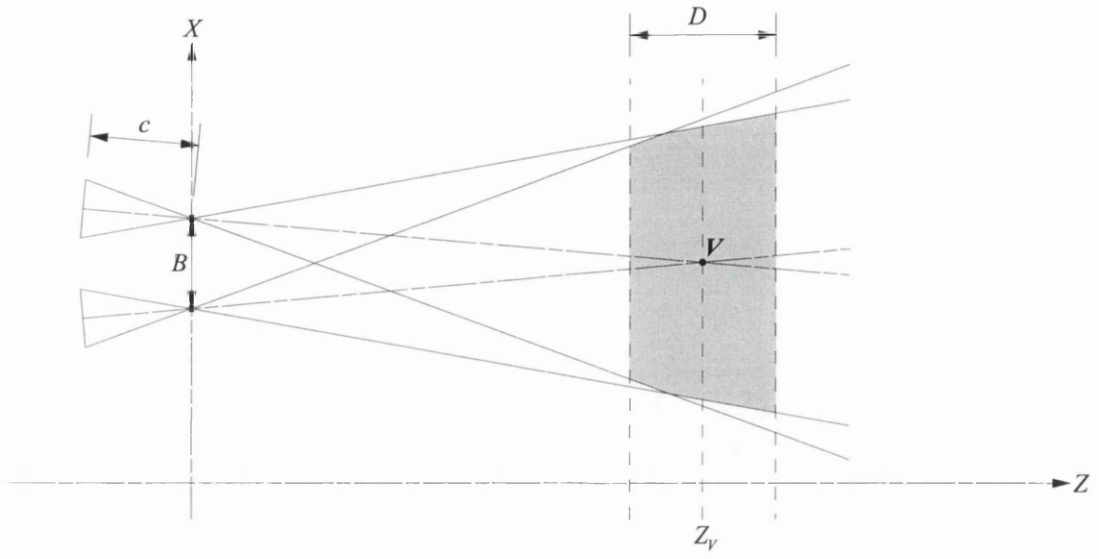


Figure 5.3. A similar viewing volume achieved with a shorter object distance and a shorter principal distance (cf. figure 5.1).

centred at the central vergence point. Consider the viewing volume obtained for a vergence point at the back of the working volume, i.e. 0.5m farther away from the cameras than the central vergence point. The width and height of the resulting viewing volume will now be larger due to the expansion of the fields of view of the cameras. The converse occurs for a vergence point at the front of the working volume with the obtained viewing volume becoming reduced in width and height.

The expansion or reduction in the dimensions of the viewing volume for different vergence points will cause associated changes in the image resolution, depth resolution and disparity limits. The change in dimensions of the viewing volume and hence the magnitude of these effects is proportional to the angular field of view of the cameras. It would clearly be beneficial to reduce these effects as much as possible. This could be achieved by choosing to adopt a long principal distance and long object distance and thereby achieve the required narrow field of view. In fact, these effects would be completely removed as principal distance and object distance tend to infinity, i.e. through orthographic, rather than perspective, projection!

Another important factor favouring the choice of lenses with long principal distances is that such lenses generally exhibit less lens distortion than similar lenses with a shorter principal distance. Therefore, long principal distance lenses will generally provide higher accuracy than short principal distance lenses.

It was decided that 5.0m would be the maximum object distance that could be accommodated within the office type of environment that the Active Stereo Probe would be situated. 5.0m was therefore adopted as the distance to the centre of the working volume for the ASP stereo vision system. According to the working volume requirements specification, the vergence points would then be constrained to lie within the depth range 4.5m to 5.5m and the total working volume would extend from 4.25m to 5.75m in depth.

5.4. Determination of principal distance

5.4.1. Objective

The objective of this experiment was to determine a suitable principal distance for the lenses to be used in the ASP stereo vision system. The suitability of a given principal distance would be judged by its ability to produce the required viewing volume throughout the specified working volume. The viewing volume specified in the performance requirements for the ASP stereo vision system was a cuboid of approximate dimensions 1.0m wide × 1.0m high × 0.5m deep centred on the vergence point. The vergence point was in turn specified to lie within a 1.0m cube. In the previous section, the central vergence point, i.e. the centre of this cube, had been chosen to lie 5.0m from the cyclopean origin of the stereo system.

5.4.2. Method

A two stage method was adopted for this analysis. The first stage was to investigate the dimensions of the viewing volume produced at the central vergence point for a range of lens principal distances. This was done using the ASP sensor head model as follows.

All of the intrinsic parameters in the model, except principal distance, were set up with appropriate values derived from the camera manufacturer's technical literature [Sony, 1990]. The extrinsic parameters were set up with the azimuth actuator origin, $({}^{W}X_{A}, {}^{W}Y_{A}, {}^{W}Z_{A})$, set to $(150\text{mm}, 0, 0)$ and $(-150\text{mm}, 0, 0)$ for camera 1 and camera 2 respectively, thus providing the initial trial stereo baseline of 300mm. The cameras were then verged on the central vergence point, (0.0m,0.0m,5.0m) in world co-ordinates. All other translations were set to zero, as were all rotations except for the azimuth actuator rotation, Φ_{A}, required to bring the cameras to convergence. This is shown in the part of the simulation log shown in figure 5.4.

```
Intrinsic Parameters
--------------------

Camera 1: Sony XC77CE Camera with 50mm lens
Camera 2: Sony XC77CE Camera with 50mm lens

                              Camera 1          Camera 2

Principal distance (mm):                +50               +50
Pixel Height (um):              +10.89            +10.89
Pixel Width (um):               +10.78            +10.78
Number of Rows:                    576               576
Number of Columns:                 768               768

Extrinsic Parameters
--------------------

Camera 1: Default Extrinsic Parameters
Camera 2: Default Extrinsic Parameters

                              Camera 1          Camera 2

Azimuth X (mm):                   -150              +150
Azimuth Y (mm):                     +0                +0
Azimuth Z (mm):                     +0                +0
Azimuth Rotation (Y):          -1.71837          +1.71837
Azimuth Wobble (Y):                 +0                +0
Azimuth Wobble (X):                 +0                +0

Elevation X (mm):                   +0                +0
Elevation Y (mm):                   +0                +0
Elevation Z (mm):                   +0                +0
Elevation Rotation (Y):             +0                +0
Elevation Wobble (Y):               +0                +0
Elevation Wobble (X):               +0                +0

Camera X (mm):                      +0                +0
Camera Y (mm):                      +0                +0
Camera Z (mm):                      +0                +0
Camera Rotation (Y):                +0                +0
Camera Wobble (Y):                  +0                +0
Camera Wobble (X):                  +0                +0

Camera 1 Vergence Point: (+0, +0, +5)
Camera 2 Vergence Point: (+0, +0, +5)
```

Figure 5.4. Part of the simulation log for the working volume experiments showing the values of the intrinsic and extrinsic parameters.

The principal distance in the model was then set to the particular value under investigation and the four corner points of each of the verged cameras' image planes were back-projected onto the world co-ordinate planes ${}^{W}Z = 4.75\text{m}$ and ${}^{W}Z = 5.25\text{m}$ (i.e. the planes forming the front and back of the required 0.5m deep viewing volume). The resulting width and height of the viewing volume were then determined from the world co-ordinates of the back-projected image plane corner points. This process, which is illustrated in plan view in figure 5.5, was repeated for a number of trial principal distances including the 50mm given in the initial configuration.

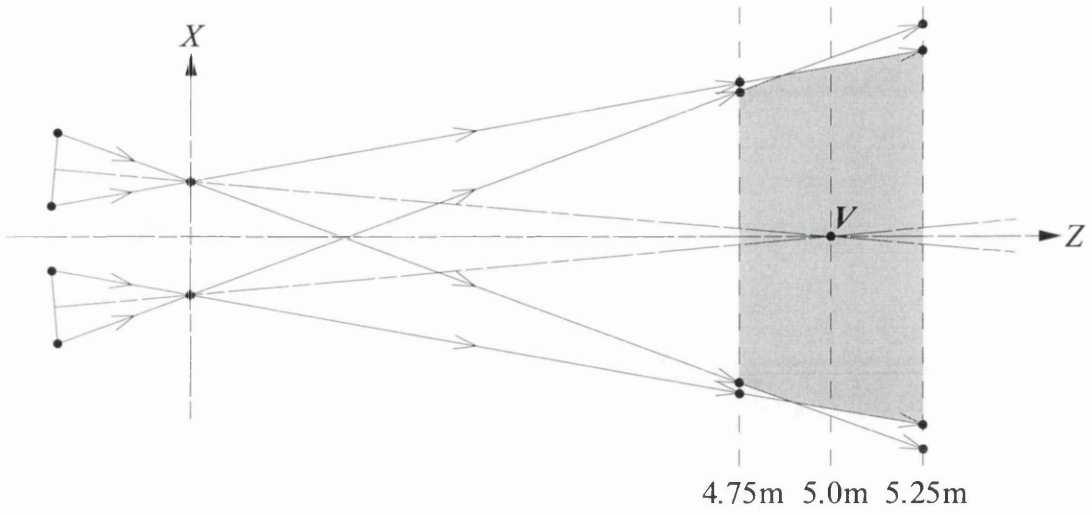


Figure 5.5. Determination of the working volume by back projecting the image plane corner points onto the world co-ordinate planes ${}^{W}Z = 4.75\text{m}$ *and* ${}^{W}Z = 5.25\text{m}$.

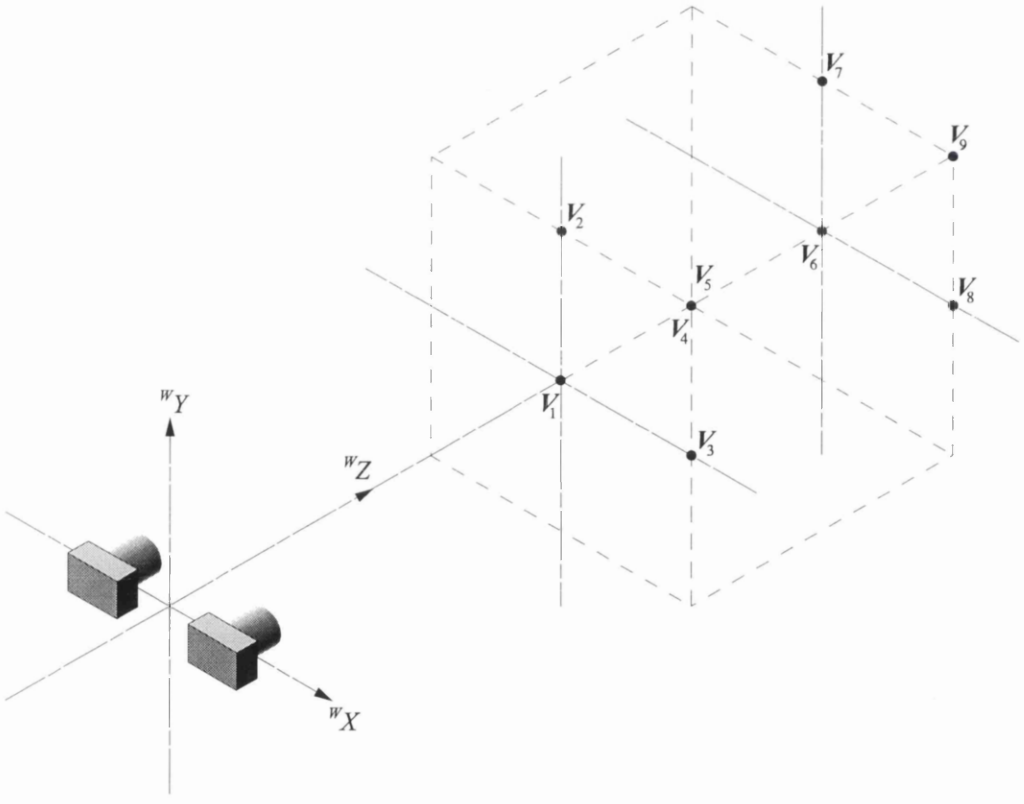


Figure 5.6. The nine trial vergence points, $\mathbf{V}_1, \mathbf{V}_2, \cdots, \mathbf{V}_9$, *used during the experiments.*

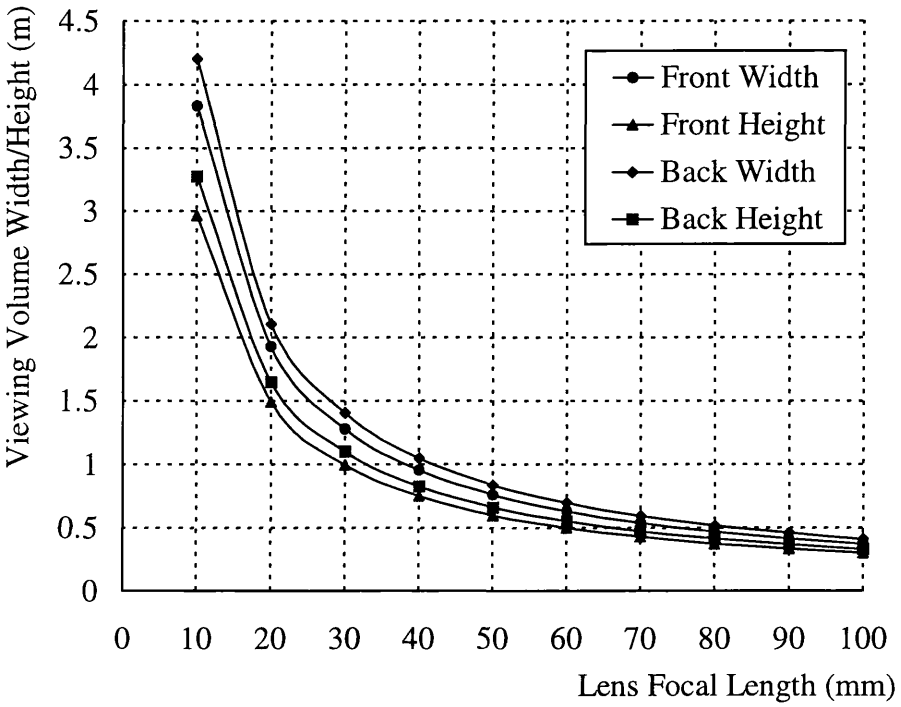


Figure 5.7. Graph of viewing volume width and height at the front and back planes ($^{W}Z = 4.75$m and $^{W}Z = 5.25$m respectively) vs. lens principal distance when the cameras are verged to (0.0, 0.0, 5.0m).

The principal distance found to provide a viewing volume most closely fitting the specification at the central vergence point was chosen for closer investigation in the second stage of the analysis. The objective of this analysis was to determine the suitability of the viewing volume derived from the chosen principal distance for vergence points at the extremes of the working volume.

Nine vergence points, $V_1, V_2, \cdots, V_9$, at the extremes of the working volume, as illustrated in figure 5.6, were chosen for this analysis: those with world co-ordinates $(0, 0, 4.5\text{m})$, $(0.5, 0, 4.5\text{m})$, $(0, 0.5, 4.5\text{m})$, $(0.5, 0.5, 4.5\text{m})$, $(0, 0, 5.0\text{m})$, $(0, 0, 5.5\text{m})$, $(0.5, 0, 5.5\text{m})$, $(0, 0.5, 5.5\text{m})$ and $(0.5, 0.5, 5.5\text{m})$. Due to the symmetry of the working volume about the ^{W}X- and ^{W}Y-axes, this selection of vergence points was acceptable for testing the viewing volume for the entire working volume.

5.4.3. Working volume vs. principal distance results

Figure 5.7 shows the graph of viewing volume against principal distance obtained for the central vergence point, $(0, 0, 5.0\text{m})$. This graph clearly shows that the viewing volume is not square in the X-Y plane and that it expands from the front plane to the

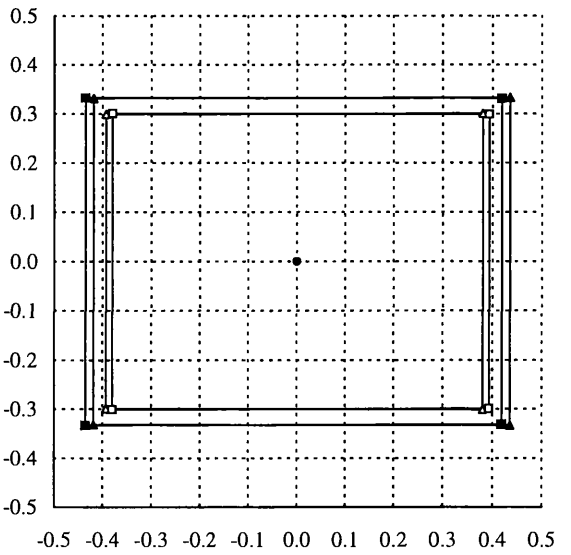


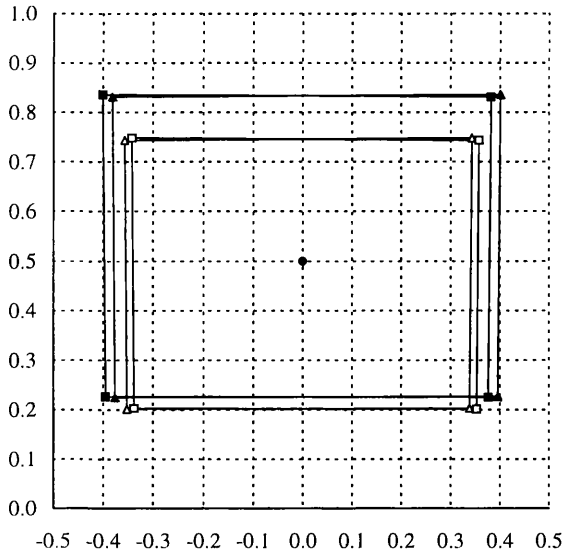
Figure 5.8. Viewing volume obtained for the central vergence point, V_5, $(0, 0, 5.0\text{m})$.

back plane. Primarily, these effects are due respectively to the aspect ratio of the CCD sensor and the expansion of the cameras' fields of view with distance. Consequently, the required viewing volume of 1.0m × 1.0m × 0.5m cannot be obtained exactly.

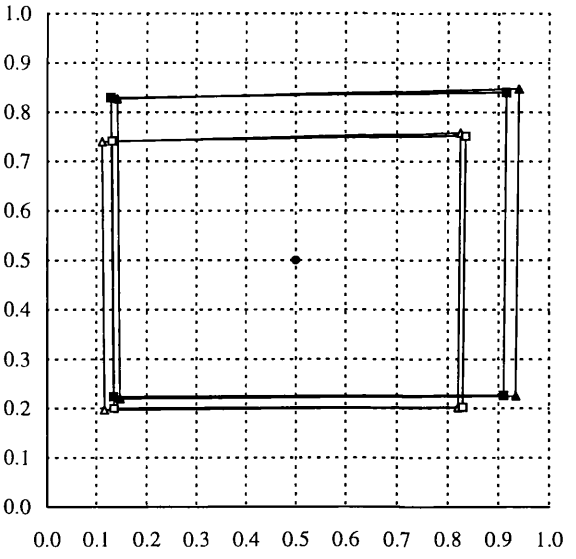
The closest compromise solution would result from the selection of a principal distance of around 35mm. This principal distance would provide a viewing volume that would be slightly wider than the required width of 1.0m but also slightly shorter than the required height of 1.0m. Unfortunately Sony miniature lenses were only available in 16mm, 25mm and 50mm focal lengths. The viewing volumes provided by principal distances of 50mm and 25mm were respectively smaller and larger than required by approximately equal amounts. However, 50mm was chosen in preference to 25mm because, as described in section 5.2, the larger principal distance would provide the more consistent size of viewing volume across the entire working volume, as shown in the next section.

5.4.4. Working volume vs. vergence point results

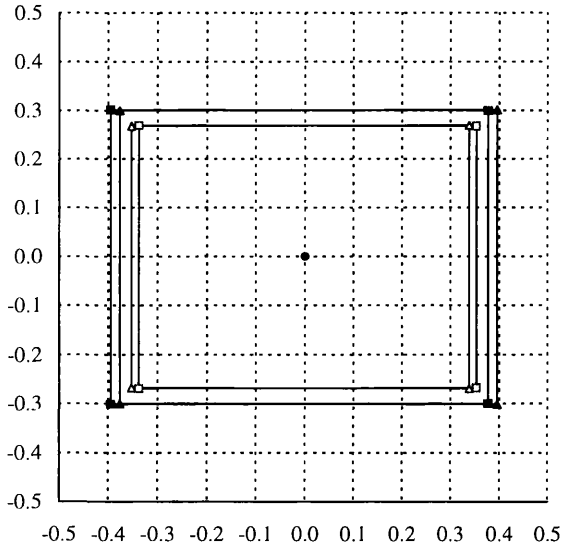
Figure 5.8, figure 5.9 and figure 5.10 plot the viewing volumes obtained for all nine of the trial vergence points and the chosen principal distance of 50mm. In each of the plots the X and Y axes correspond to the ${}^{W}X$ and ${}^{W}Y$ axes respectively; the circular point is the vergence point; the open and closed triangular points are the projections of the left camera sensor's corner points onto the front and back planes of the viewing



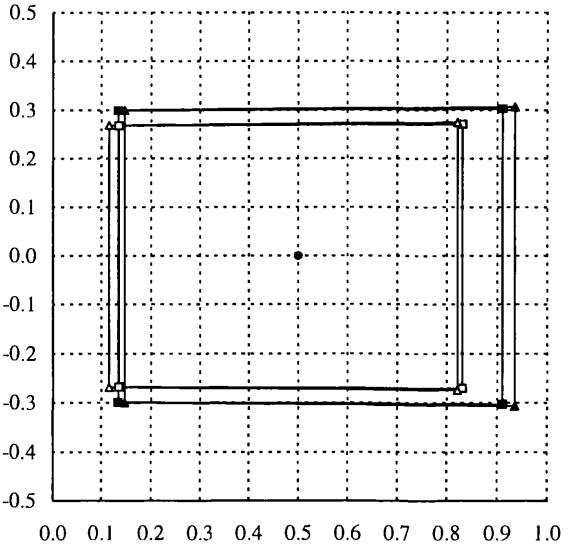
Vergence Point V_2, (0, 0.5, 4.5)



Vergence Point V_4, (0.5, 0.5, 4.5)



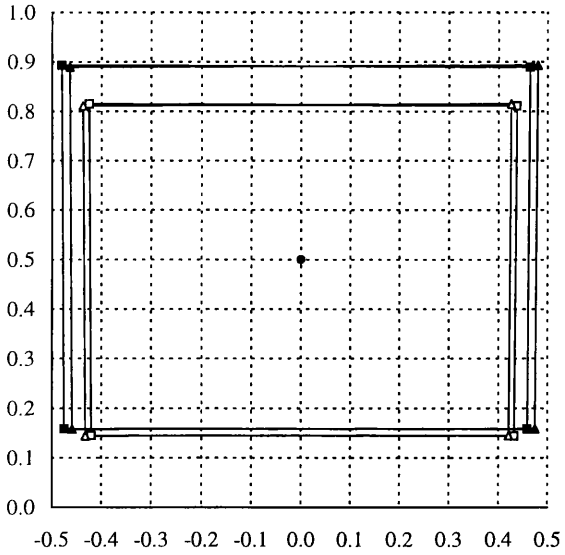
Vergence Point V_1, (0, 0, 4.5)



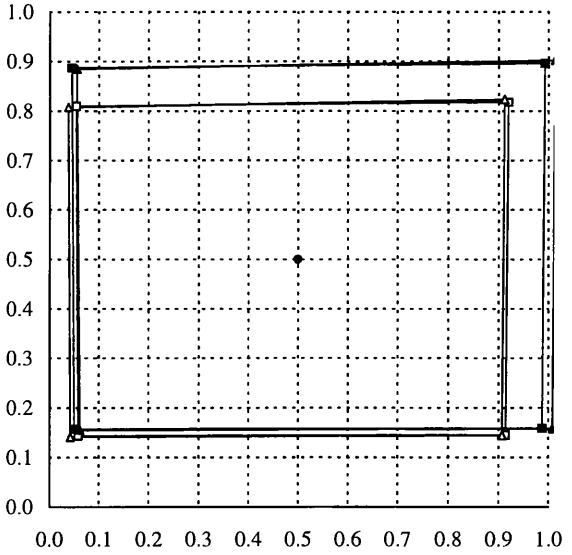
Vergence Point V_3, (0.5, 0, 4.5)

Figure 5.9. Viewing volume obtained for the vergence points (0, 0, 4.5m), (0.5, 0, 4.5m), (0, 0.5, 4.5m) *and* (0.5, 0.5, 4.5m).

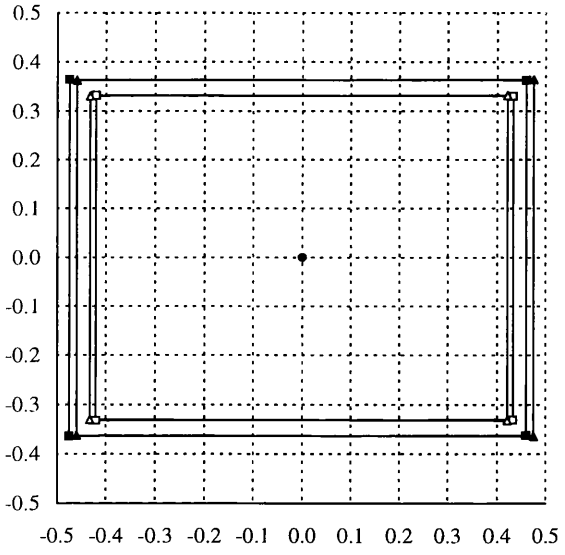
volume respectively; and the open and closed square points are the respective corresponding points for the right camera.



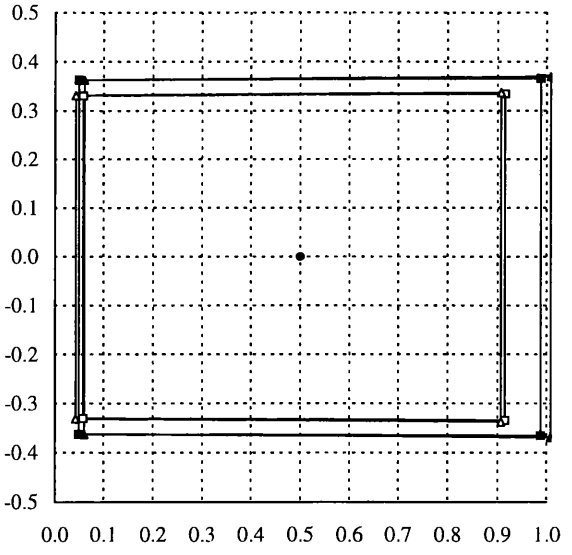
Vergence Point V_7, (0, 0.5, 5.5)



Vergence Point V_9, (0.5, 0.5, 5.5)



Vergence Point V_6, (0, 0, 5.5)



Vergence Point V_8, (0.5, 0, 5.5)

Figure 5.10. Viewing volume obtained for the vergence points (0, 0, 5.5m), (0.5, 0, 5.5m), (0, 0.5, 5.5m) *and* (0.5, 0.5, 5.5m).

5.4.5. Interpretation of results

Of the range of principal distances available for Sony miniature lenses, 50mm was found to provide the closest viewing volume to that specified for the ASP stereo vision system for the central vergence point (0.0, 0.0, 5.0m). It was therefore decided that the viewing volume provided by this principal distance be investigated at vergence points at the extremes of the viewing volume to ensure its suitability for the ASP stereo vision system.

For all four of the trial vergence points at the front of the working volume, i.e. with ${}^{W}Z = 4.5\text{m}$, the achievable viewing volume was found to be approximately 0.7m wide × 0.5m high for the required depth of 0.5m. The achievable viewing volume for the four trial vergence points at the back of the working volume, i.e. with ${}^{W}Z = 5.5\text{m}$, was found to be approximately 0.85m wide × 0.65m high for the required depth of 0.5m. It was preferable to obtain a viewing volume that was smaller than the required 1.0m × 1.0m × 0.5m than one that was larger by the same amount. Therefore a principal distance of 50mm was concluded to provide the best compromise working volume for a mean object distance of 5.0m.

5.5. Determination of stereo baseline

5.5.1. Objective

Having chosen the cameras, the mean object distance and principal distance for the ASP stereo vision system, the only parameter of the basic stereo imaging configuration remaining to be selected was the camera baseline separation. The objective of the experiment described in this section was therefore to determine a suitable baseline separation for the cameras in the ASP stereo vision system.

As the baseline separation of a stereo vision system is increased so the disparity between the stereo image points of observed world points also increases. Since stereo matching algorithms are limited in their operating range of disparity, increasing the baseline will reduce the volume of world points whose image points can be matched, and hence whose position can be measured. Conversely, the finite disparity resolution of stereo matching algorithms means that increasing the baseline will result in more precise measurements of the positions of those remaining world points whose stereo image points can be matched.

A horizontal disparity limit of ±20 pixels was specified for the MSSM stereo matching algorithm because of the enormous amount of computation required to

process larger disparities. The first objective in choosing the baseline for the ASP stereo vision system was therefore to ensure that the range of disparities resulting from all possible world points within the required working volume is as close as possible to the ±20 pixels limit. This would ensure that it would be possible to measure the positions of all world points occurring within the required working volume with the maximum possible resolution.

In addition to the horizontal disparity limit of ±20 pixels, a vertical disparity limit of ±0.5 pixels was also specified for the MSSM stereo matching algorithm. This was done to ensure that a one dimensional search would be sufficient during stereo matching since the two dimensional search that would be required for larger vertical disparities vastly increased the amount of computation required by MSSM. Therefore a second objective of the experiments described in this section was to ensure that the vertical range of disparities resulting from all points within the required working volume did not exceed the vertical disparity limit of ±0.5 pixels.

5.5.2. Theory

The *absolute maximum disparity* and the *maximum range of disparities* for a particular stereo imaging configuration and viewing volume are the two main quantities that will be investigated in the experiments described later in this section. However, this sub-section first attempts to describe what these quantities mean and how they can be derived.

In the stereo imaging configuration illustrated in figure 5.11, consider a point $\boldsymbol{p}_{i,0}$ on the image plane of the left camera. When $\boldsymbol{p}_{i,0}$ is back-projected, it intersects the front plane of the viewing volume at the world point $\boldsymbol{P}_{i,1}$. Subsequent projection of $\boldsymbol{P}_{i,1}$ onto the image plane of the right camera results in the right image point $\boldsymbol{p}_{i,1}$. If the image co-ordinates of $\boldsymbol{p}_{i,0}$ and $\boldsymbol{p}_{i,1}$ are respectively $(x_{i,0}, y_{i,0})$ and $(x_{i,1}, y_{i,1})$ (where the x co-ordinate is the column position and the y co-ordinate is the row position of an image point), then the disparity between $\boldsymbol{p}_{i,0}$ and $\boldsymbol{p}_{i,1}$ for that particular stereo configuration is simply:

$$\delta_{x_{i,1}} = x_{i,1} - x_{i,0} \tag{5.3}$$

$$\delta_{y_{i,1}} = y_{i,1} - y_{i,0} \tag{5.4}$$

where $\delta_{x_{i,1}}$ and $\delta_{y_{i,1}}$ are, respectively, the horizontal and vertical components of disparity resulting from the stereo observation of the world point $\boldsymbol{P}_{i,1}$.

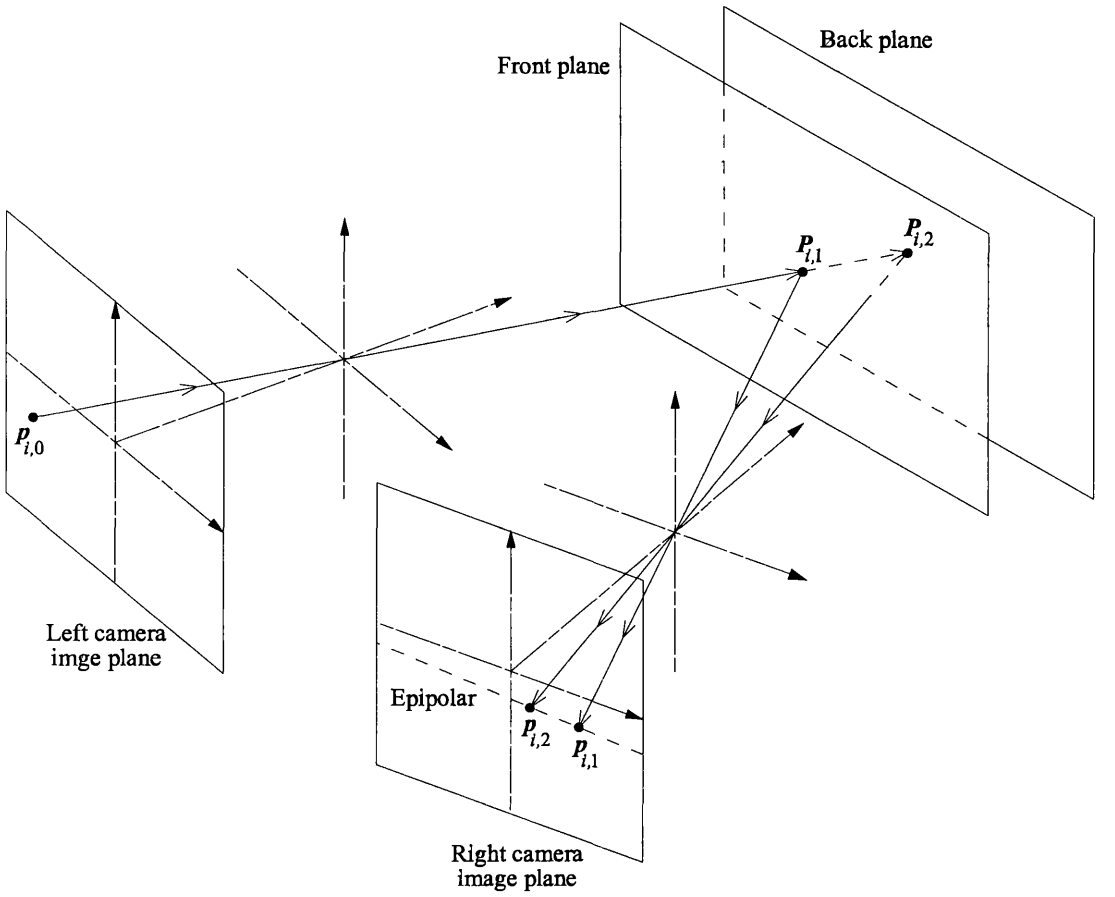


Figure 5.11. Derivation of a right image epipolar line from a left image point $\boldsymbol{p}_{i,0}$. See text for details of this process.

Similarly, if $\boldsymbol{p}_{i,0}$ is back-projected onto the back plane of the viewing volume, instead of the front-plane, the result is the world point $\boldsymbol{P}_{i,2}$. When $\boldsymbol{P}_{i,2}$ is in turn projected onto the right image plane, the result is the right image point $\boldsymbol{p}_{i,2}$ with co-ordinates $(x_{i,2}, y_{i,2})$. The horizontal and vertical components of disparity resulting from the stereo observation of $\boldsymbol{P}_{i,2}$ are then:

$$\delta_{x_{i,2}} = x_{i,2} - x_{i,0} \tag{5.5}$$

$$\delta_{y_{i,2}} = y_{i,2} - y_{i,0} \tag{5.6}$$

By definition, the straight line passing through $\boldsymbol{p}_{i,1}$ and $\boldsymbol{p}_{i,2}$ is the right image epipolar line corresponding to the left image point $\boldsymbol{p}_{i,0}$. Also, since $\boldsymbol{p}_{i,1}$ and $\boldsymbol{p}_{i,2}$ are the right image projections of the points $\boldsymbol{P}_{i,1}$ on the front plane and $\boldsymbol{P}_{i,2}$. on the back plane of the viewing volume respectively, only the portion of the epipolar line between $\boldsymbol{p}_{i,1}$ and $\boldsymbol{p}_{i,2}$ corresponds to world points within the viewing volume. Image points on the epipolar line to either side of $\boldsymbol{p}_{i,1}$ and $\boldsymbol{p}_{i,2}$ correspond to world points outside the viewing

volume. Consequently, the maximum possible disparity for $\boldsymbol{p}_{i,0}$ resulting from the stereo observation of points within the viewing volume must be the disparity between $\boldsymbol{p}_{i,0}$ and either $\boldsymbol{p}_{i,1}$ or $\boldsymbol{p}_{i,2}$. Therefore, if there are N possible points (i.e. pixels), $\mathrm{p}_{i,0}, 0 \le i < N$, in the left image, then the *absolute maximum disparity* that can possibly occur for that particular stereo imaging configuration and viewing volume is:

$$\delta_{x,\max} = \max\left|x_{i,j} - x_{i,0}\right|, \quad i = 0,1,\cdots,N\text{-}1,\ j = 1,2 \tag{5.7}$$

$$\delta_{y,\max} = \max\left|y_{i,j} - y_{i,0}\right|, \quad i = 0,1,\cdots,N\text{-}1,\ j = 1,2 \tag{5.8}$$

where $\delta_{x,\max}$ and $\delta_{y,\max}$ are respectively the absolute maximum horizontal and vertical disparity.

Similarly, the range of possible disparities for the left image point $\boldsymbol{p}_{i,0}$ resulting from world points within the viewing volume is simply the difference between the disparities at $\boldsymbol{p}_{i,1}$ and $\boldsymbol{p}_{i,2}$. Therefore the range of possible horizontal and vertical disparities for the left image point $\boldsymbol{p}_{i,0}$ are respectively:

$$\Delta_{x_i} = \left|\delta_{x_{i,2}} - \delta_{x_{i,1}}\right| = \left|\left(x_{i,2} - x_{i,0}\right) - \left(x_{i,1} - x_{i,0}\right)\right| \tag{5.9}$$

$$\Delta_{y_i} = \left|\delta_{y_{i,2}} - \delta_{y_{i,1}}\right| = \left|\left(y_{i,2} - y_{i,0}\right) - \left(y_{i,1} - y_{i,0}\right)\right| \tag{5.10}$$

The *maximum range of disparities* for all left image points $\mathrm{p}_{i,0}, 0 \le i < N$ is then simply:

$$\Delta_{x,\max} = \max\left|\left(x_{i,2} - x_{i,0}\right) - \left(x_{i,1} - x_{i,0}\right)\right|, \quad i = 0,1,\cdots,N-1 \tag{5.11}$$

$$\Delta_{y,\max} = \max\left|\left(y_{i,2} - y_{i,0}\right) - \left(y_{i,1} - y_{i,0}\right)\right|, \quad i = 0,1,\cdots,N-1 \tag{5.12}$$

where $\Delta_{x,\max}$ and $\Delta_{y,\max}$ are respectively the maximum range of horizontal and vertical disparities.

5.5.3. Method

In order to determine the most suitable baseline separation of the cameras in the ASP stereo vision system, a two stage approach to the investigation was once again adopted. The aim of the first stage of experimentation was to characterise the disparity limit imposed by a range of stereo baselines but for only a single vergence point: the central vergence point. The configuration that was thus found to impose a disparity limit closest to that specified for the MSSM stereo matching algorithm for the central

viewing volume was chosen for more detailed investigation in the second stage of experiments. The aim of the second stage was then to characterise the disparity limit imposed by the selected baseline for each viewing volume formed by verging the cameras to the nine previously described trial vergence points. This would therefore verify that the disparity limit imposed by the selected baseline was within the specification for the MSSM stereo matching algorithm for the entire working volume.

In both stages of the investigation, the disparity limit imposed by a particular stereo imaging configuration was obtained by using the ASP stereo head model to determine the *absolute maximum disparity* and the *maximum range of disparities* for that imaging configuration. The method described in the previous sub-section was used; i.e. that of projecting a number of left image points onto the front and back planes of the viewing volume and then onto the right image plane. However it was deemed unnecessary and too time consuming to calculate the absolute maximum disparity and maximum range of disparity for all N (i.e. $575 \times 756 = 434700$) left image points as indicated in equations 5.7, 5.8, 5.11 and 5.12.

Instead only five points, $\boldsymbol{p}_{i,0}, i = 1,2,\cdots,5$, at the centre and four corners of the left image plane and five, $\boldsymbol{p}_{i,0}, i = 6,7,\cdots,10$, at the equivalent locations on the right image plane were used in the analysis. Note that in the case of the right image points, $\boldsymbol{p}_{i,0}, i = 6,7,\cdots,10$, the two image points, $\boldsymbol{p}_{i,1}$ and $\boldsymbol{p}_{i,2}$, derived for each are the end points of the epipolar line in the left image plane. The pixel co-ordinates of all ten of the original image points, $\boldsymbol{p}_{i,0}, i = 1,2,\cdots,10$, are given in table 5.2.

Left Image Point	Right Image Point	Column Position (x co-ordinate)	Row Position (y co-ordinate)
$\boldsymbol{p}_{1,0}$	$\boldsymbol{p}_{6,0}$	377	290
$\boldsymbol{p}_{2,0}$	$\boldsymbol{p}_{7,0}$	-377	290
$\boldsymbol{p}_{3,0}$	$\boldsymbol{p}_{8,0}$	0	0
$\boldsymbol{p}_{4,0}$	$\boldsymbol{p}_{9,0}$	377	-290
$\boldsymbol{p}_{5,0}$	$\boldsymbol{p}_{10,0}$	-377	-290

Table 5.2. The ten image points used in the disparity limits analysis.

The results of the first stage of experiments are given in sub-section 5.5.4 and the results of the second stage are given in sub-section 5.5.5.

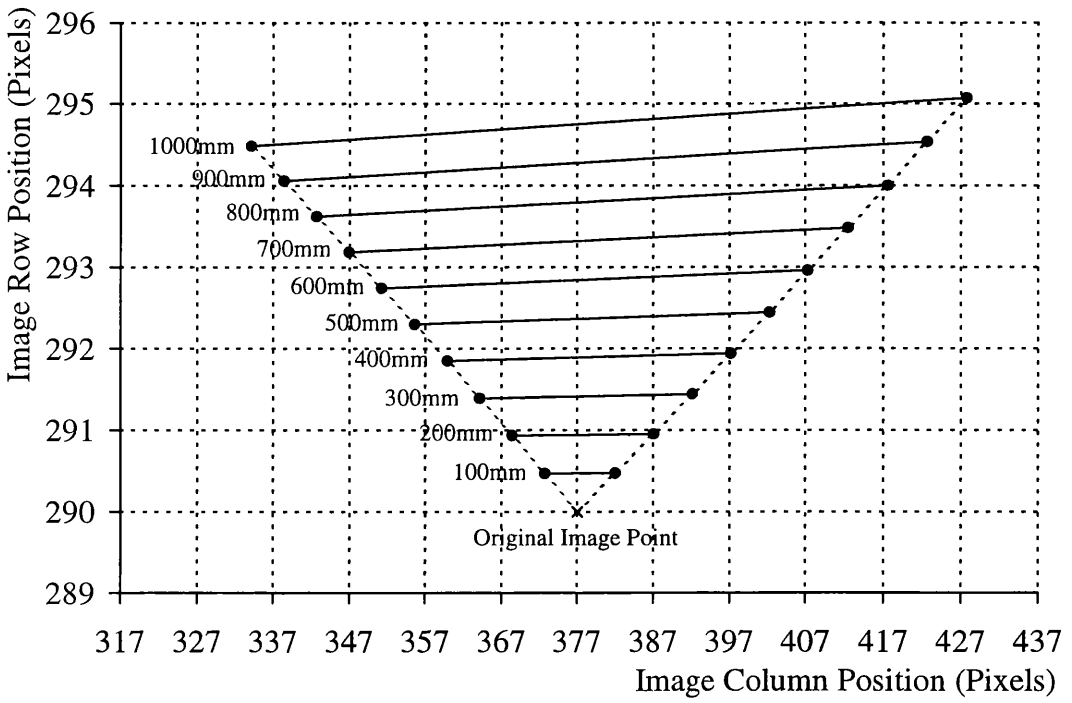


Figure 5.12. Plot of right image epipolars derived for the original left image point $\boldsymbol{p}1_0$, (377, 290), *with the cameras verged to the central vergence point,* (0.0, 0.0, 5.0m), *and camera baselines ranging from 100mm to 1000mm.*

5.5.4. Disparity limits vs. baseline results

The objective of this experiment was to determine the disparity limit imposed by each of a range of baselines for the viewing volume at the centre of the working volume specified for the ASP stereo vision system. This was done by setting up the ASP stereo head model with a range of stereo baselines from 100mm to 1000mm and verging the cameras to the central vergence point, (0.0, 0.0, 5.0m). For each of these stereo imaging configurations, the method described in the previous section was used to obtain plots of the right image epipolars corresponding to the left image points $\boldsymbol{p}n_0, n = 1,2,\cdots,5$. Similar plots of the left image epipolars were derived for each of the right image points $\boldsymbol{p}n_0, n = 6,7,\cdots,10$. An example of such a plot, for the right image epipolars derived from the left image point $\boldsymbol{p}1_0$, is shown in figure 5.12.

The absolute maximum disparity and the maximum range of disparity were derived for all ten of these epipolar plots. These figures are plotted in figure 5.13, which is a graph of absolute maximum horizontal disparity, $\delta_{x,\max}$, and maximum range of horizontal disparity, Δ_x, versus camera baseline, and figure 5.14, which is a graph of absolute maximum vertical disparity, $\delta_{y,\max}$, and maximum range of vertical disparity, Δ_y, versus camera baseline.

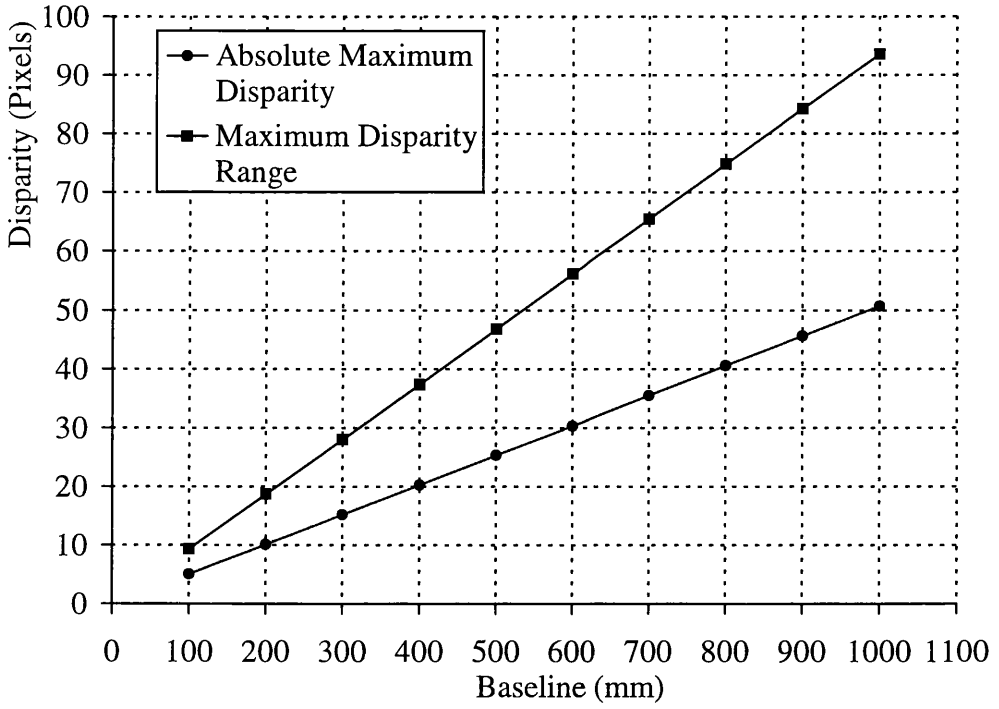


Figure 5.13. Graph of absolute maximum horizontal disparity, $\delta_{x,\max}$*, and maximum range of horizontal disparity,* Δ_x*, versus camera baseline for the central vergence point,* (0.0, 0.0, 5.0m).

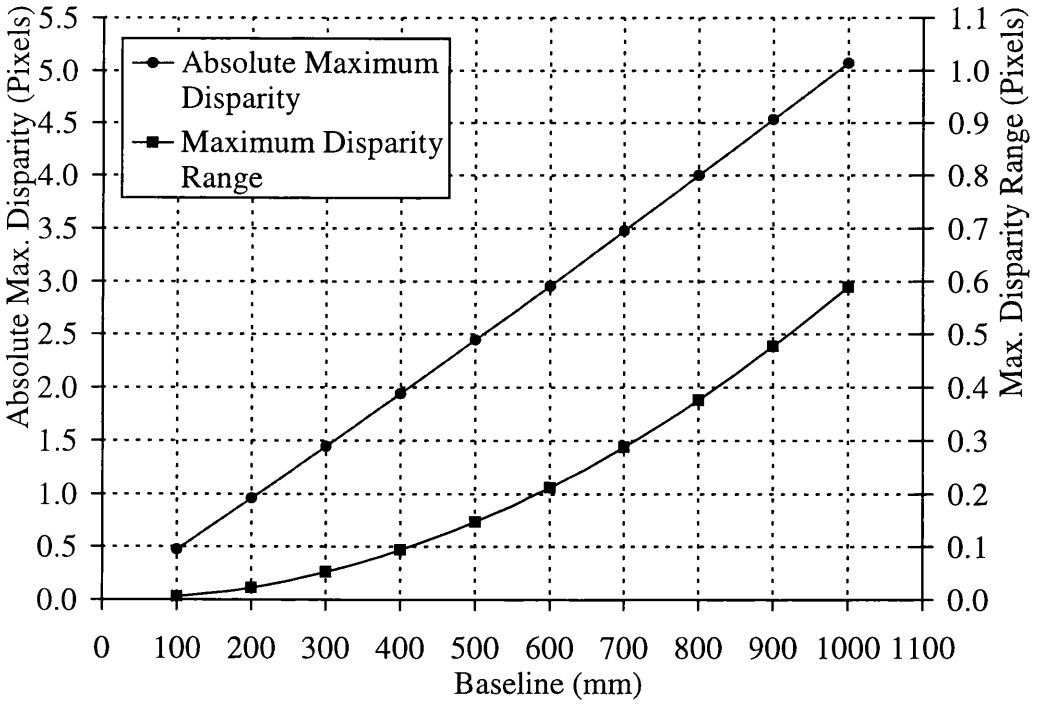


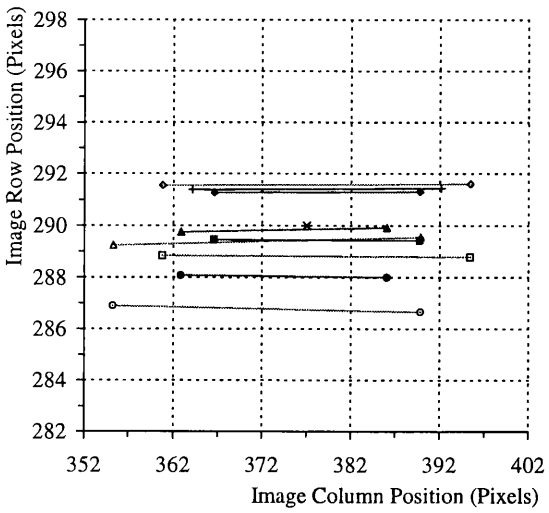
Figure 5.14. Graph of absolute maximum vertical disparity, $\delta_{y,\max}$*, and maximum range of vertical disparity,* Δ_y*, versus camera baseline for the central vergence point,* (0.0, 0.0, 5.0m).

5.5.5. Disparity limits vs. vergence point results

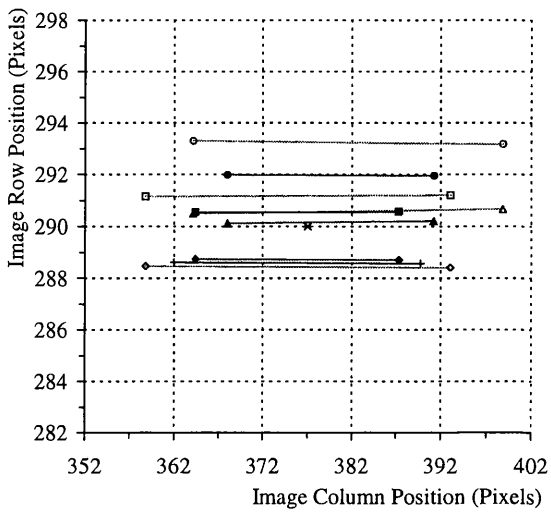
From the results presented in sub-section 5.5.4 and other factors 300mm was chosen as a suitable baseline for further investigation in the second stage of experiments. (See the conclusions in the following section for a discussion of the reasons for making this selection). The objective of the second stage of experiments was then to characterise the disparity limit required by this baseline value of 300mm for all nine of the trial vergence points. The results of this second stage of experiments are presented in this sub-section.

Each graph in figure 5.15 shows the epipolar lines obtained in the opposite image for one of the ten image points, $\boldsymbol{pn}_0, n = 1,2,\cdots,10$. The epipolar lines obtained for all of the nine trial vergence points have been plotted for each image point. The graphs on the left hand side in figure 5.15 were generated from an original image point on the left image plane (i.e. one of $\boldsymbol{pn}_0, n = 1,2,\cdots,5$) and hence show the epipolar lines that result in the right image. The graphs on the right hand side show the left image epipolars resulting from an original image point on the right image plane (i.e. one of $\boldsymbol{pn}_0, n = 6,7,\cdots,10$).

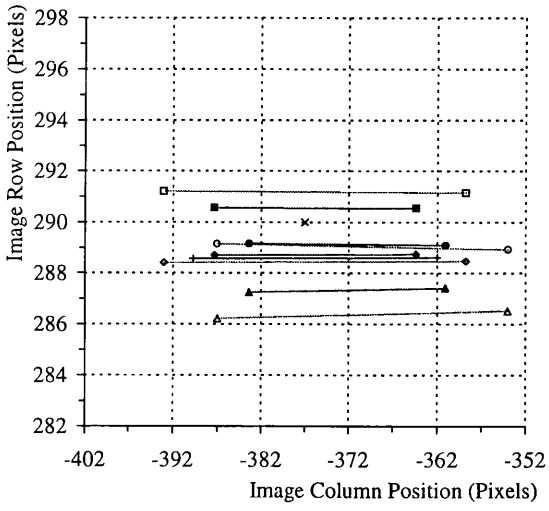
The epipolar plots of figure 5.15 are summarised in table 5.3, which gives the absolute maximum horizontal and vertical disparities, $\delta_{x,\max}$ and $\delta_{y,\max}$ respectively, and maximum range of horizontal and vertical disparity, Δ_x and Δ_y respectively, for each of the trial vergence points.



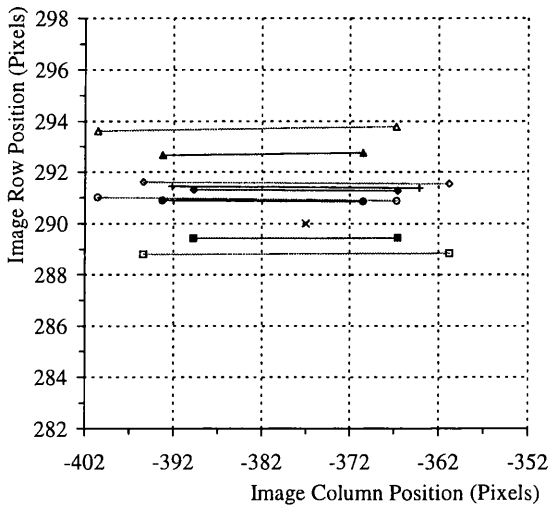
Right image epipolars derived from the left image point $\boldsymbol{p}1_0$, (377, 290).



Left image epipolars derived from the right image point $\boldsymbol{p}6_0$, (377, 290).



Right image epipolars derived from the left image point $\boldsymbol{p}2_0$, (-377, 290).

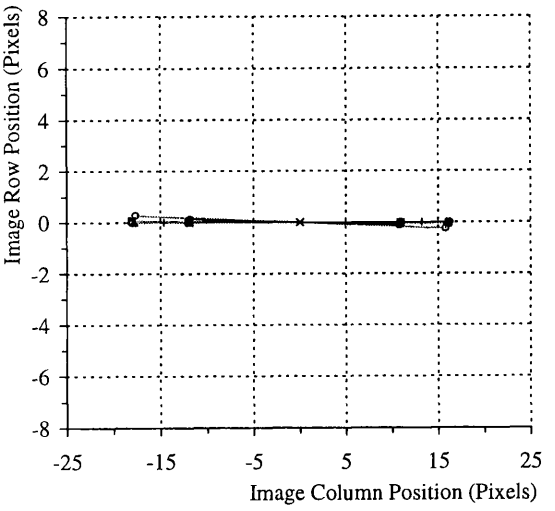


Left image epipolars derived from the right image point $\boldsymbol{p}7_0$, (-377, 290).

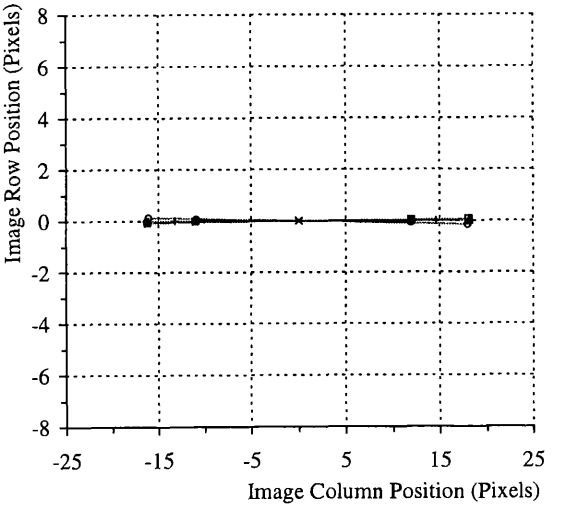
Legend

× Original Image Point	(0.0, 0.0, 5.0) Epipolar
(0.0, 0.0, 4.5) Epipolar	(0.5, 0.0, 4.5) Epipolar
(0.0, 0.5, 4.5) Epipolar	(0.5, 0.5, 4.5) Epipolar
(0.0, 0.0, 5.5) Epipolar	(0.5, 0.0, 5.5) Epipolar
(0.0, 0.5, 5.5) Epipolar	(0.5 ,0.5, 5.5) Epipolar

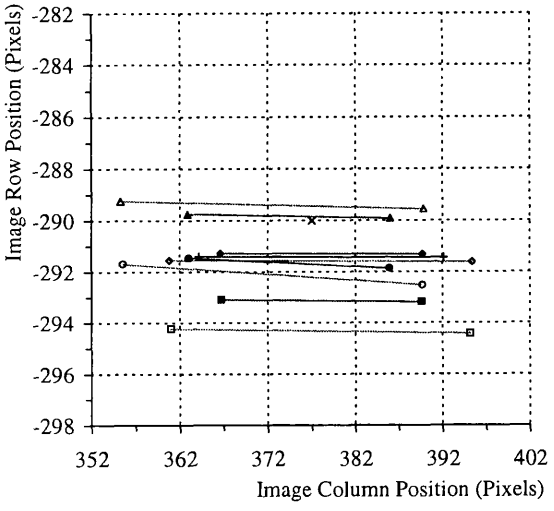
Figure 5.15. Epipolar lines resulting in the opposite image from one of the image points, $\boldsymbol{p}n_0, n = 1,2,\cdots,10$, *for each of the nine different vergence points as indicated in the legend.*



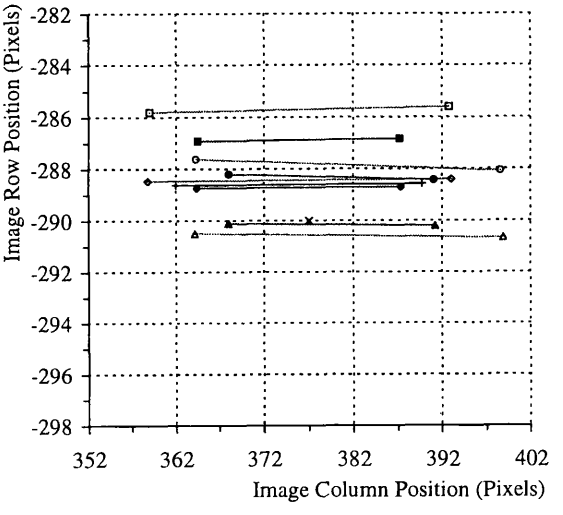
Right image epipolars derived from the left image point $\boldsymbol{p}3_0$, (0, 0).



Left image epipolars derived from the right image point $\boldsymbol{p}8_0$, (0, 0).



Right image epipolars derived from the left image point $\boldsymbol{p}4_0$, (377, -290).

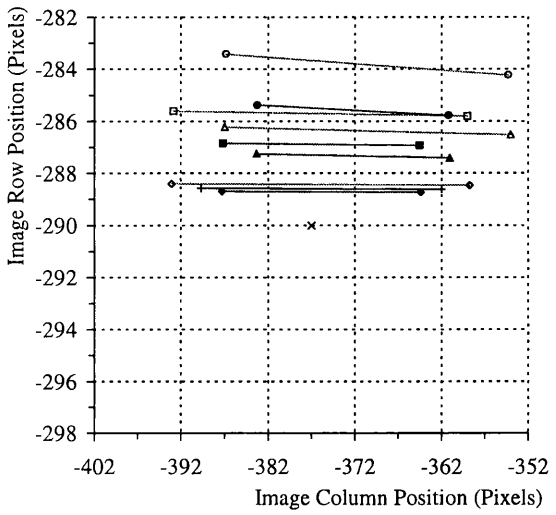


Left image epipolars derived from the right image point $\boldsymbol{p}9_0$, (377, -290).

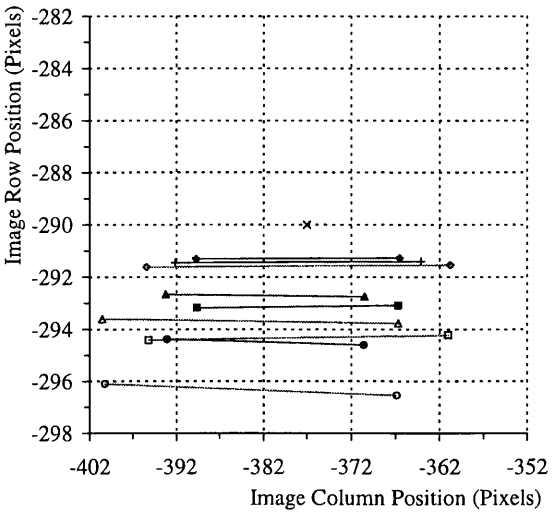
Legend

× Original Image Point	(0.0, 0.0, 5.0) Epipolar
(0.0, 0.0, 4.5) Epipolar	(0.5, 0.0, 4.5) Epipolar
(0.0, 0.5, 4.5) Epipolar	(0.5, 0.5, 4.5) Epipolar
(0.0, 0.0, 5.5) Epipolar	(0.5, 0.0, 5.5) Epipolar
(0.0, 0.5, 5.5) Epipolar	(0.5 ,0.5, 5.5) Epipolar

Figure 5.15 (continued).



Right image epipolars derived from the left image point $\mathbf{p}5_0$, (-377, -290).



Left image epipolars derived from the right image point $\mathbf{p}10_0$, (-377, -290).

Figure 5.15 (continued).

Vergence Point	Horizontal Disparity (Pixels)		Vertical Disparity (Pixels)	
	Absolute Max.	Max. Range	Absolute Max.	Max. Range
(0.0, 0.0, 4.5)	18.401	34.613	1.614	0.072
(0.5, 0.0, 4.5)	23.564	34.696	3.78	0.307
(0.0, 0.5, 4.5)	18.201	34.641	4.386	0.195
(0.5, 0.5, 4.5)	23.568	34.727	6.572	0.831
(0.0, 0.0, 5.0)	15.154	28.011	1.446	0.052
(0.0, 0.0, 5.5)	12.720	23.132	1.310	0.039
(0.5, 0.0, 5.5)	16.239	23.193	2.759	0.169
(0.0, 0.5, 5.5)	12.729	23.169	3.158	0.095
(0.5, 0.5, 5.5)	16.255	23.231	4.616	0.407

Table 5.3. Absolute maximum disparity and maximum range of disparity (horizontal and vertical) for a baseline of 300mm and each of the trial vergence points.

5.5.6. Interpretation of results

5.5.6.1. Horizontal Disparity Limit

It can be seen from figure 5.13 that an absolute maximum horizontal disparity of approximately 15 pixels results when the cameras are verged to (0.0, 0.0, 5.0m) with a baseline of 300mm. Therefore a disparity limit of at least ±15 pixels would be required to guarantee complete matching of the stereograms captured using this stereo camera configuration. The performance requirements for the ASP stereo vision system stated that the horizontal disparity limit should be between ±10 pixels and ±20 pixels throughout the working volume. Therefore the disparity limit of ±15 pixels predicted for the central vergence point lies approximately midway within the range of acceptable values.

Furthermore, larger disparity values were expected to result from vergence points closer than (0.0, 0.0, 5.0m) and smaller disparity values were expected from farther vergence points. It therefore seemed reasonable to conclude that a baseline of 300mm would be most likely to meet the disparity limit requirements for all vergence points within the working volume. Hence a baseline of 300mm was chosen for closer investigation in the second stage of experimentation in which the disparity limits required by a range of vergence points was investigated.

The horizontal disparity results obtained from the second stage of experimentation, as summarised in table 5.3, support the choice of 300mm for the stereo camera baseline in the ASP stereo vision system. It can be seen from this table that the absolute maximum horizontal disparity ranges from 12.720 pixels for the vergence point (0.0, 0.0, 5.5m) to 23.568 pixels for the vergence point (0.5, 0.5, 4.5m). This value of 23.568 pixels initially seems to imply that a horizontal disparity limit of at least ±23.568 pixels would be required to ensure the complete matching of stereograms captured when the cameras were verged to (0.5, 0.5, 4.5m). This would be outside the performance requirements specified for the ASP stereo vision system.

Close examination of the epipolar plots obtained for the vergence point (0.5, 0.5, 4.5m) shows that it is the horizontal disparity between the right image point $\boldsymbol{p}7_0$ and the end point of the left image epipolar, $\boldsymbol{p}7_1$, that is -23.568 pixels. However the disparity between $\boldsymbol{p}7_0$ and the other end point of the epipolar, $\boldsymbol{p}7_2$, is only 10.281 pixels. Therefore the total range of horizontal disparity for $\boldsymbol{p}7_0$ is 33.849 pixels. If the stereo matching process were to be offset to the centre of this epipolar line, then a horizontal disparity limit of only ±16.925 pixels (half of the total range) would be required to ensure complete matching for $\boldsymbol{p}7_0$.

In practice, the ASP stereo vision system will be accurately calibrated. This will make it possible to predict the actual locations of the epipolar lines in captured stereo pairs of images prior to stereo matching (using precisely the same method as used to perform the analysis above). Hence, the offsets required to restrict the disparity search process for each pixel to only that part of the epipolar line corresponding to the viewing volume could also be calculated. If these offsets were to be applied to the matching process, then the horizontal disparity limit requirements of the ASP stereo vision system would be satisfactorily met by a baseline of 300mm. This can be seen from table 5.3, which shows that the maximum range of horizontal disparity ranges from 23.132 pixels, for the vergence point (0.0, 0.0, 5.5m), to 34.727 pixels, for the vergence point (0.5, 0.5, 4.5m). The required disparity limit would therefore range from ±11.566 pixels to ±17.364 pixels: well within the specified range of ±10 pixels to ±20 pixels.

5.5.6.2. Vertical Disparity Limit

The performance requirements for the ASP stereo vision system stated that the vertical disparity limit should be no greater than ±0.5 pixels. Figure 5.14 shows that, in terms of the absolute maximum disparity range, only baselines of 100mm or less are capable of achieving this level of performance for the vergence point (0.0, 0.0, 5.0m). From table 5.3 it can be seen that, for a baseline of 300mm which was deemed to be the most suitable for meeting the horizontal disparity limit requirements, the absolute maximum vertical disparity is greater than 1.0 pixels for all vergence points and is as high as 6.572 pixels for the vergence point (0.5, 0.5, 4.5m). It was therefore clear that the vertical disparity limit requirement could not be met in terms of absolute maximum vertical disparity.

The solution was again to consider offsetting the stereo matching process so that only the range of disparity is significant. Figure 5.14 shows that, for the vergence point (0.0, 0.0, 5.0m), all of the baselines tested result in a range of vertical disparity that is less than the 1.0 pixels required to meet the specified limit of ±0.5 pixels. Referring to table 5.3, it can be seen that the range of vertical disparity is less than the 1.0 pixels for all of the trial vergence points when the baseline is 300mm.

5.6. Conclusions

5.6.1. Review of objectives

The specific objective of the experiments described in this chapter was to establish the basic configuration of the ASP sensor head. This comprised the selection of suitable

cameras and lenses and the determination of an appropriate combination of camera baseline and mean object distance. The suitability of a particular configuration would be judged by its ability to both image the working volume specified for the ASP system, and acquire stereo images with a range of disparities that fills, but does not exceed, the dynamic range of the MSSM stereo matcher.

5.6.2. Summary of results

5.6.2.1. Choice of cameras

The selection of the cameras for the ASP system was an essential preliminary stage to the experiments described in this chapter because the imaging characteristics of the cameras would have a fundamental influence on the overall performance of the stereo vision system. Sony XC77RR-CE cameras were chosen primarily for their small size and high resolution. The electronic shutter facility provided by these cameras was an additional factor in their favour.

5.6.2.2. Choice of mean object distance

The combination of as large a mean object distance and as long a principal distance as possible was deemed most likely to result in a viewing volume that would meet the specified requirements for the entire working volume. 5.0m was adopted as the mean object distance for the ASP stereo vision system since this was considered to be the maximum that could be accommodated within the office type of environment that was envisaged for the system. Vergence points would hence be constrained to lie within the depth range 4.5m to 5.5m and the total working volume would extend from 4.25m to 5.75m in depth.

5.6.2.3. Determination of lens principal distance

The criterion used for the selection of a suitable lens principal distance for the ASP stereo vision system was the dimensions of the viewing volume that would be obtained for any vergence point within the working volume. After extensive simulation experiments using the ASP stereo head model, 50mm was found to be the most suitable principal distance available for Sony lightweight miniature lenses. The resulting viewing volume was found to range in size from approximately 0.7m wide $\times$ 0.5m high $\times$ 0.5m deep at the front of the working volume to approximately 0.85m wide $\times$ 0.65m high $\times$ 0.5m deep at the back.

5.6.2.4. Determination of stereo baseline

The stereo baseline for the ASP stereo vision system was selected on the basis of the total possible range of disparity to occur in stereograms captured using a particular stereo configuration. After performing a series of simulation experiments using the ASP stereo head model, it was found that the specified requirements could not be met in terms of the absolute maximum disparity. However it was found that the requirements could be met if the stereo matching process was offset to the centre of the epipolar line for each image point to be matched. In this case, a stereo baseline of 300mm was found to provide a satisfactory level of performance when the whole working volume is taken into account.

5.6.3. Conclusions

The experiments described in this chapter have successfully achieved the specific objective of determining the basic configuration of the ASP sensor head. However, they have also been successful in the wider, and perhaps more important, context of demonstrating the power of the simulation based methodology for designing stereo vision systems.

The experiments showed that it was not possible to meet exactly the desired specification for the ASP sensor system due to the restricted range of components available. However, the adopted simulation based design methodology allowed the range of configurations that are possible in practice to be explored. It was then possible to select the configuration that was both achievable in practice and most closely met the desired specification for the ASP sensor system.

In addition to allowing a specific configuration for the ASP sensor system to be selected, the experiments also provided a characterisation of the expected level of performance that would be provided by the selected configuration. This would facilitate the task of testing the completed ASP sensor head and validate its correct manufacture and operation.

Having performed the analysis once to design the ASP sensor head, the analysis could be repeated easily to achieve any of the following:

1. redesign the ASP sensor system to take account of new components as they become available;

2. explore the performance achieved by alternative configurations;

3. explore the sensitivity of the system to changes in the specification of any of the components;

4. to design a completely new stereo vision system with a different specification.

Chapter 6

Sensor system actuation

6.1. Introduction

6.1.1. Background

As outlined in Chapter 3, the dynamic calibration scheme proposed for the ASP stereo vision system is intended to be able to make use of feedback from the sensor actuators. After every camera movement, it is proposed that actuator positional feedback will be used to derive new values for the external orientation parameters in the collinearity equation model of the stereo vision system. However it is clear that any inaccuracy in the positional feedback will result in an error in the derived estimates of the external orientation parameters and hence error between the "actual" position of surface points and the "measured" positions derived from subsequent stereo surface recovery.

It is therefore a necessary part of the design process for the ASP sensor head to specify the level of positional feedback accuracy required to achieve the desired level of surface measurement accuracy. However, prior to this work, the relationship between the sensor position error and the surface measurement error had not be analysed in any great detail. Indeed it was not even known what order of magnitude of positional error would have a significant effect on surface measurements. Consequently it was not possible initially to specify the sensor actuation requirements for the ASP stereo vision system.

6.1.2. Objective

The objective of the experiments described in this chapter was to remedy the above situation by performing an in-depth analysis of the relationship between sensor position inaccuracies and surface measurement error in the context of the ASP stereo vision system. Armed with this information it would then be possible to specify fully the sensor actuation requirements of the ASP stereo vision system.

This objective was to be achieved by performing a series of simulation experiments using the ASP sensor system model. A clear picture of the relationship between each separate component of the sensor position error and the resulting surface measurement error was to be obtained first. This would allow the requirements for a system incorporating all of the components of sensor position error to achieve a given level of surface measurement accuracy to be investigated. The performance requirements for a sensor head suitable for use in the proposed dynamic calibration scheme could then be determined. Finally, a sensor head could be designed and fabricated to meet the specified requirements.

6.1.3. Parameters under analysis

The orientation of each camera in the ASP sensor head is to be controlled by two rotational actuators: one each for the azimuth and elevation axes. In the proposed dynamic calibration scheme, the external orientation parameters of each camera are to be derived from feedback of the measured actuator rotation angles. Clearly, any inaccuracy in this positional feedback will result in errors in the derived external orientation parameters and ultimately in surface measurement error. Errors in the rotation angles of the actuators are therefore one set of parameters that require to be investigated during this analysis. The azimuth actuator and elevation actuator rotation angle are represented by the parameters Φ_A and Ω_E respectively in the ASP sensor head model.

According to the data provided by actuator manufacturers [Physik Instrumente, 1990; Micro-Controle, 1990; Ealing 1990], the only other significant source of pointing error introduced by high quality rotational actuators is "off-axis wobble". This is an unavoidable deviation of the rotation of the actuator from the nominal plane of rotation. It was therefore deemed necessary to also consider the effects of this "off-axis wobble" during the analysis. In the ASP sensor head model, two parameters are used to represent "off-axis wobble" for each actuator: ϕ_A and ω_A for the azimuth actuator and ω_E and κ_E for the elevation actuator. See Chapter 4 for a more detailed description of these parameters.

Actuator angle error and "off-axis wobble" will exhibit random and systematic components, both of which will vary dynamically during operation of the ASP stereo vision system. However additional errors will be introduced into the derived estimates of the external orientation parameters due to the errors in the assumed positions of the actuators and cameras that will result from unavoidable construction errors. These errors will remain static during dynamic system operation and it is therefore

conceivable that they may be "calibrated out" to a certain extent. Nevertheless it was decided that their effects should be considered during the analysis to ascertain whether normal construction tolerances alone would be sufficient to allow the required level of surface measurement accuracy. The parameters of the ASP sensor head model associated with these static errors are: the azimuth actuator position, ${}^{W}X_{A}$, ${}^{W}Y_{A}$ and ${}^{W}Z_{A}$, the elevation actuator position, ${}^{A}X_{E}$, ${}^{A}Y_{E}$ and ${}^{A}Z_{E}$, and the camera position, ${}^{E}X_{C}$, ${}^{E}Y_{C}$ and ${}^{E}Z_{C}$. Also considered was the possibility of errors in the orientation of the camera which were represented by the three camera rotation parameters: K_{C}, κ_{C} and ϕ_{c}.

The internal orientation parameters of the collinearity equation model were not considered relevant to the analysis reported in this chapter since it was assumed that they will remain constant during dynamic operation of the ASP stereo vision system. In any case, the internal orientation parameters cannot be derived directly from the actuator feedback. Instead there are many reported methods that could be used for deriving them during a static calibration stage prior to dynamic operation of the ASP stereo vision system. In the event that the internal orientation parameters do drift during dynamic system operation, the proposed closed-loop extension to dynamic calibration that uses bundle adjustment could be used to refine their estimated values.

To summarise, the effects of 36 parameters of the ASP stereo system model on surface measurement accuracy were to be investigated during the analysis. The 36 parameters to be investigated were the 18 external parameters for each camera: ${}^{W}X_{A}$, ${}^{W}Y_{A}$, ${}^{W}Z_{A}$, Φ_{A}, ϕ_{A}, ω_{A}, ${}^{A}X_{E}$, ${}^{A}Y_{E}$, ${}^{A}Z_{E}$, Ω_{E}, ω_{E}, κ_{E}, ${}^{E}X_{C}$, ${}^{E}Y_{C}$, ${}^{E}Z_{C}$, K_{C}, κ_{C} and ϕ_{c}. The effects of errors in the internal orientation parameters and the effects of stereo-matching errors were not considered, since these were deemed to be the concern of static calibration rather than dynamic calibration.

6.1.4. Absolute error and residual error

During these investigations into surface measurement error it was important that the difference between absolute and residual error in the measurement of a surface was noted. The *absolute error* is simply the point-wise difference between the actual surface and the measured surface when both are defined with respect to the same world co-ordinate frame. This representation of measurement error fails to take into account the fact that the global error between the positions (and possibly the orientations) of the actual and measured surfaces is often relatively large. Once this global error has been removed from the measured surface, the point-wise residual, or *residual error*, is often significantly smaller (see figure 6.1 and figure 6.2). For most

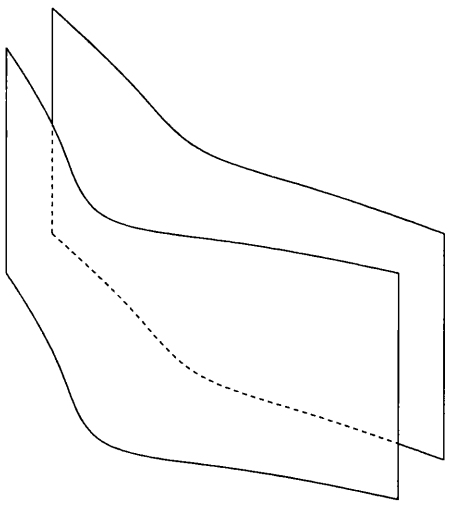


Figure 6.1. Two surfaces before removal of global errors.

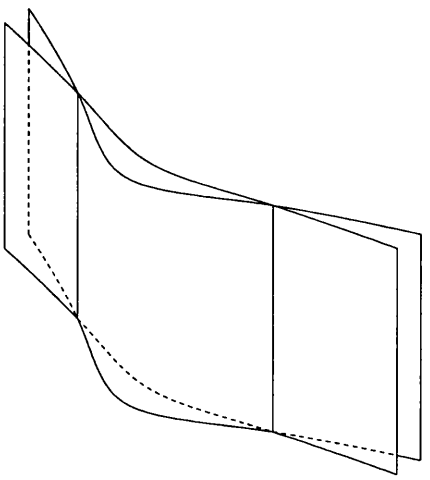


Figure 6.2. Two surfaces after removal of global errors.

applications, the residual error is the more informative since it represents the actual error in recovering the shape and dimensions of the surface itself.

6.1.5. Basic approach

The experiments described in this chapter were carried out in two stages. The aim of the first set of experiments that are described in section 6.3 was to determine the individual relationship between surface measurement accuracy and a representative sub-set of the sensor position parameters listed in sub-section 6.1.3. The results of these experiments formed the basis of the second set of experiments that analysed the effect on surface measurement accuracy of errors in all of the sensor position parameters acting simultaneously. This second set of experiments is described in section 6.4. From these results, the positional feedback accuracy and construction tolerances required to meet a given specification of surface measurement could be obtained.

6.2. Basic analytical procedure

A common basic analytical procedure was employed for all of the experiments described in this chapter. This procedure involved the generation of two sets of virtual world co-ordinate points within the ASP sensor system model. The first set of virtual world co-ordinate points represent the "actual" position of points on a 3D surface that is to be measured. The second set of virtual world co-ordinate points represent the "measured" positions of the first set of points obtained by simulated stereo surface

recovery. The difference between the locations of the "actual" and "measured" virtual world co-ordinate points represents the error in the measurement process.

The four steps in this common analytical procedure can be summarised as follows:

1. The ASP sensor system model is set-up in a particular configuration of the stereo cameras. This configuration simulates the "actual" position of the cameras used to acquire a stereo pair of images. With the cameras in this position, the first set of "actual" virtual world points and the corresponding set of virtual stereo image points are generated inside the ASP sensor system model.

2. The sensor position parameters of the ASP sensor system model's cameras were then altered to simulate either errors in the positional feedback provided by the actuators or errors in the construction of the sensor head.

3. The stereo pairs of virtual image points that were generated in Step 1 were back-projected through the altered cameras to form a second set of virtual world co-ordinate points. These new virtual world co-ordinate points represent the "measured" locations of the "actual" world points generated in Step 1.

4. The surface measurement error induced by altering the sensor position parameters was calculated from the differences between the co-ordinates of the "actual" and "measured" virtual world co-ordinate points.

Each of the above four steps of the common analytical procedure are described in greater detail in the following sub-sections.

6.2.1. Step 1: generation of "actual" surface points

Step 1 of the basic procedure for simulating surface measurement error was the generation of the first set of virtual world co-ordinate points within the ASP sensor system model. This first set of virtual world co-ordinate points represent the "actual" position of points on a 3D surface that is to be measured. Ten "actual" virtual world points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$, illustrated in figure 6.3, were chosen for the analysis. These points, at the extremes of the viewing volume, were chosen since they were deemed to be the most likely to generate the maximum residual error for a given vergence point. Sub-section 6.2.3 more fully explains this selection of world co-ordinate points.

The first stage of the generation of the ten "actual" virtual world points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$, was to define ten image points, $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{10}$, on the image plane of Camera One within the ASP sensor system model. $\mathbf{p}_1$ was defined to lie at the centre of the image

plane and $\mathbf{p}_2$, $\mathbf{p}_3$, $\mathbf{p}_4$ and $\mathbf{p}_5$ were defined to lie at its four corners. $\mathbf{p}_6,\mathbf{p}_7,\cdots,\mathbf{p}_{10}$, were defined to lie at exactly the same locations as $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_5$ respectively. The reason for this seemingly redundant second set of definitions will shortly become apparent. The image co-ordinates of $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{10}$ are given in table 6.1.

Camera One Image Point	Column Position (pixels)	Row Position (pixels)	Plane of Projection
$\mathbf{p}_1$	0	0	Back
$\mathbf{p}_2$	-384	-288	Back
$\mathbf{p}_3$	384	-288	Back
$\mathbf{p}_4$	-384	288	Back
$\mathbf{p}_5$	384	288	Back
$\mathbf{p}_6$	0	0	Front
$\mathbf{p}_7$	-384	-288	Front
$\mathbf{p}_8$	384	-288	Front
$\mathbf{p}_9$	-384	288	Front
$\mathbf{p}_{10}$	384	288	Front

Table 6.1. Image co-ordinates of the ten Camera One virtual image points used to generate the "actual" world points. Plane of projection refers to whether the point is projected onto the front or back of the viewing volume.

After defining the ten Camera One image points, $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{10}$, the parameters of the stereo cameras in the ASP sensor system model were set up in the initial configuration required for each individual experiment, with the cameras verged to the desired virtual vergence point, **V**. (The initial configuration for each experiment is detailed in the appropriate section describing that experiment). The first five Camera One virtual image points, $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_5$, were then back projected, using the ASP sensor system model, onto the world plane forming the back of the viewing volume (e.g. for the vergence point (0, 0, 5.0m), this was the plane ${}^{W}Z = 5.25\text{m}$). These back-projections resulted in the first five "actual" virtual world points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_5$ respectively, lying on the back plane of the viewing volume. The second five Camera One image points, $\mathbf{p}_6,\mathbf{p}_7,\cdots,\mathbf{p}_{10}$, were back projected onto the front plane of the viewing volume (${}^{W}Z = 4.75\text{m}$ in the example) to form the remaining "actual" virtual world points, $\mathbf{P}_6,\mathbf{P}_7,\cdots,\mathbf{P}_{10}$ respectively.

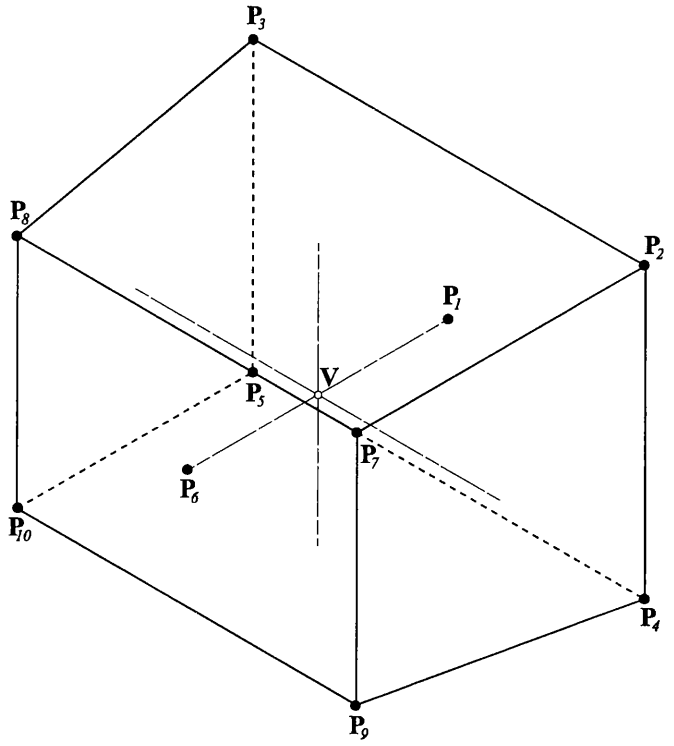


Figure 6.3. The ten "actual" virtual world co-ordinate points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$, *at the extremes of the viewing volume and the virtual vergence point,* $\mathbf{V}$.

Step 1 of the basic analytical procedure was completed by generating the ten Camera Two image points, $\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_{10}$, that corresponded to $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$ respectively. This was done simply by using the ASP sensor system model to project each of the virtual world points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$, onto the image plane of Camera Two. The whole of Step 1 thus resulted in the ten "actual" virtual world points, $\mathbf{P}_i$, $i=1,2,\cdots,10$, and the ten corresponding stereo pairs of Camera One and Camera Two image points, $\mathbf{p}_i$ and $\mathbf{q}_i$ respectively. The whole processing chain from $\mathbf{p}_4$ to $\mathbf{P}_4$ to $\mathbf{q}_4$ is illustrated in figure 6.4.

6.2.2. Step 2: introduction of parameter errors

Step 2 of the basic analytical procedure was to alter parameters in the ASP sensor system model to simulate either errors in the positional feedback provided by the actuators or errors in the construction of the sensor head. For example, consider the experiment to investigate the relationship between surface measurement error and an error of $\pm\varepsilon_{\Phi_A}$ in the positional feedback from the azimuth actuators. In this case, $\pm\varepsilon_{\Phi_A}$ is the error between the "actual" azimuth actuator rotation, Φ_A, and the "measured" value, $\hat{\Phi}_A$, provided by the actuator feedback. Thus $\hat{\Phi}_{A1}$, the "measured" value for

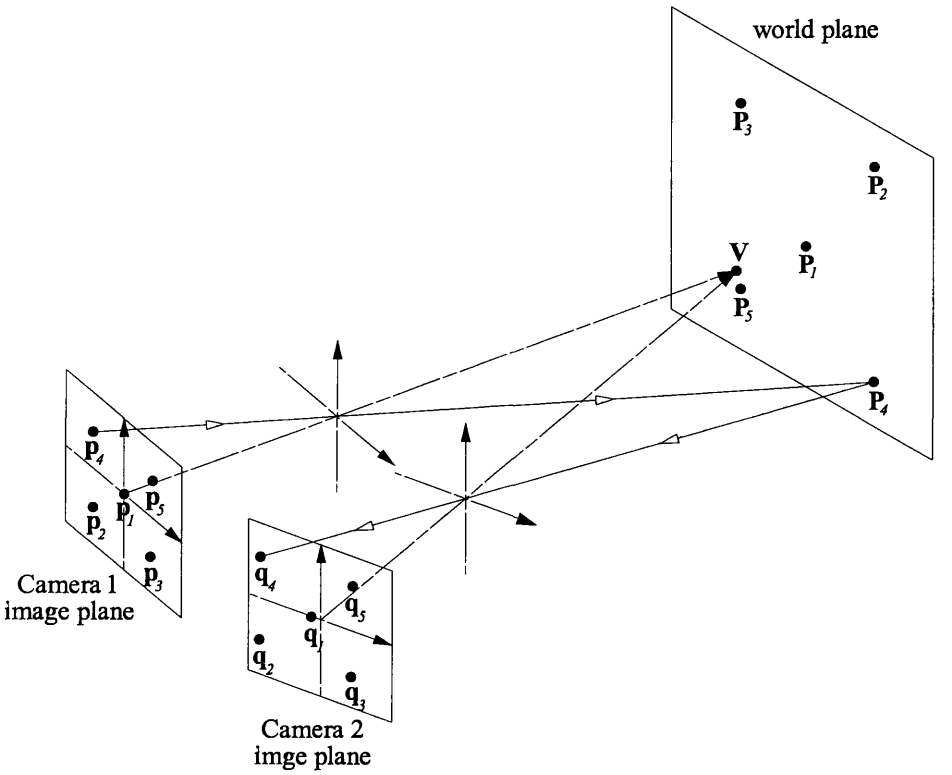


Figure 6.4. Generation of the five "actual" world points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_5$, *on the back plane of the viewing volume and their corresponding image points in Camera One,* $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_5$, *and Camera Two,* $\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_5$.

Camera One, can take either of two values, $\Phi_{A1}+\varepsilon_{\Phi_A}$ or $\Phi_{A1}-\varepsilon_{\Phi_A}$ and the "measured" value for Camera Two, $\hat{\Phi}_{A2}$, can be either $\Phi_{A2}+\varepsilon_{\Phi_A}$ or $\Phi_{A2}-\varepsilon_{\Phi_A}$, where Φ_{A1} and Φ_{A2} are the "actual" azimuth actuator rotation angles for Camera One and Camera Two respectively. Therefore, in order to find the maximum surface measurement error induced by an azimuth rotation error of $\pm\varepsilon_{\Phi_A}$, all four combinations of $\hat{\Phi}_{A1}=\Phi_{A1}\pm\varepsilon_{\Phi_A}$ and $\hat{\Phi}_{A2}=\Phi_{A2}\pm\varepsilon_{\Phi_A}$ would be applied to the ASP sensor system model.

A similar method was adopted for the introduction of errors into the elevation actuator rotation, Ω_E, camera rotation, K_C, and each component of azimuth actuator translation, elevation actuator translation and camera translation, ${}^{W}X_A$, ${}^{W}Y_A$, ${}^{W}Z_A$, ${}^{A}X_E$, ${}^{A}Y_E$, ${}^{A}Z_E$, and ${}^{E}X_C$, ${}^{E}Y_C$, ${}^{E}Z_C$ respectively.

Since two parameters, wobble angle and wobble orientation, were used to define "off-axis wobble" for each actuator or camera, a slightly different approach had to be adopted for its introduction into the model. Azimuth actuator "off-axis wobble", for

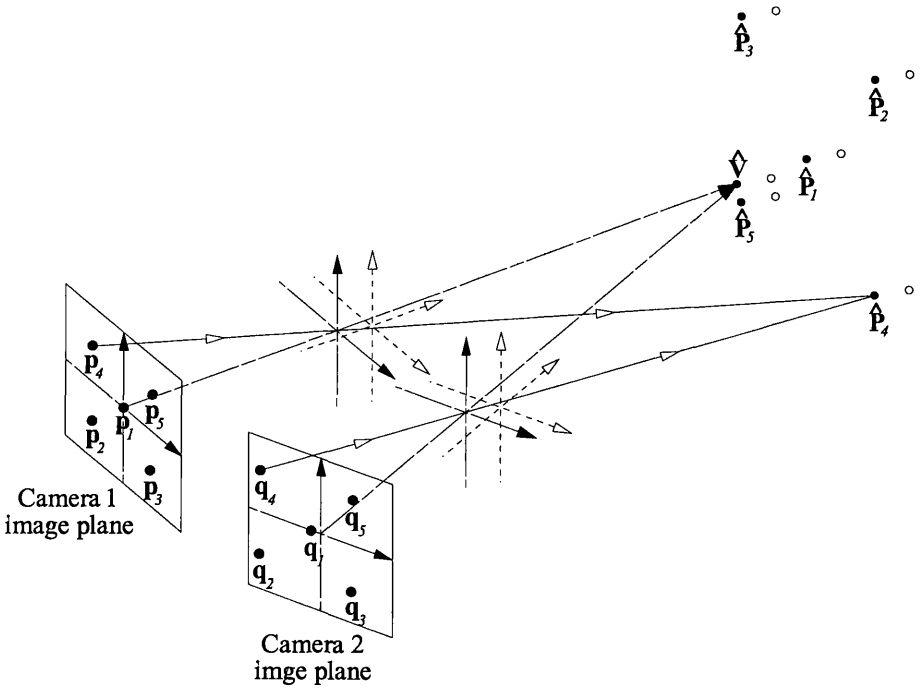


Figure 6.5. The "measured" virtual world points, $\hat{\mathbf{P}}_1,\hat{\mathbf{P}}_2,\cdots,\hat{\mathbf{P}}_5$, are generated after the introduction of parameter errors by stereo back projection of the virtual image points $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_5$ and $\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_5$.

example, was defined by a fixed parameter: wobble angle, ω_A, and a free parameter: wobble orientation, ϕ_A. Therefore, in order to find the maximum surface measurement error induced by an azimuth actuator wobble angle of ω_A, one complete revolution of wobble orientation, ϕ_A, in 10° steps, was applied independently for each camera. This resulted in a total of 1296 (36 × 36) combinations of wobble orientations for both cameras. The same procedure was used to introduce elevation actuator "off-axis wobble", κ_E and ω_E, and camera "off-axis wobble", ϕ_c and κ_C.

6.2.3. Step 3: generation of "measured" surface points

After altering the ASP sensor system model with a particular combination of parameter errors, the cameras will no longer be verged to the "actual" virtual vergence point, $\mathbf{V}$. They will instead be verged to the "measured" virtual vergence point, $\hat{\mathbf{V}}$. In Step 3 of the basic analytical procedure, the "measured" vergence point was found by stereo back projection of the image point (0,0) from both cameras after the appropriate errors had been added to the model.

Similarly, the "measured" positions of the world surface points was found by using the ASP sensor system model to stereo back project the stereo pairs of points formed

from $\mathbf{p}_1,\mathbf{p}_2,\cdots,\mathbf{p}_{10}$ and $\mathbf{q}_1,\mathbf{q}_2,\cdots,\mathbf{q}_{10}$. This resulted in ten "measured" virtual surface points, $\hat{\mathbf{P}}_1,\hat{\mathbf{P}}_2,\cdots,\hat{\mathbf{P}}_{10}$, the measured positions of the "actual" virtual world points, $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$ respectively. Figure 6.5 illustrates this stereo back-projection process for $\hat{\mathbf{P}}_1,\hat{\mathbf{P}}_2,\cdots,\hat{\mathbf{P}}_5$, the "measured" positions of $\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_5$, the five "actual" virtual world points originally on the back plane of the viewing volume.

6.2.4. Step 4: calculation of surface measurement error

The vector difference between the "actual" and "measured" virtual vergence points, $\mathbf{V}$ and $\hat{\mathbf{V}}$ respectively, induced by a particular combination of parameter errors was termed the *vergence error*, $\mathbf{E}_V$. The *maximum vergence error*, $\mathbf{E}_{V\max}$, obtained for all combinations of parameter errors in a given experiment was used as a measure of the global surface measurement error for that experiment. i.e.:

$$\mathbf{E}_{V\max} = \max_{c=0}^{c<C}\left[\hat{\mathbf{V}}_c - \mathbf{V}\right] \tag{6.1}$$

Where $\hat{\mathbf{V}}_c$ is the "measured" virtual vergence point derived for each combination, c, of parameter errors and C is the total number of combinations of parameter errors used in the experiment.

Similarly, the absolute surface measurement error for a point is simply the vector difference between the "actual" virtual world point, $\mathbf{P}$, and the corresponding "measured" virtual world point, $\hat{\mathbf{P}}$. However, in order to derive the relative surface measurement error, separate object space co-ordinate systems had to be defined for the "actual" and "measured" sets of virtual world points. The actual and measured virtual vergence points were taken as the respective origins of these co-ordinate systems. The *residual error*, $\mathbf{E}_R$, between the "actual" and "measured" virtual world points was therefore defined as:

$$\mathbf{E}_R = (\hat{\mathbf{P}} - \hat{\mathbf{V}}) - (\mathbf{P} - \mathbf{V}) \tag{6.2}$$

Note that this is simply the absolute surface measurement error at the virtual world point $\mathbf{P}$ minus the absolute error at the virtual vergence point, $\mathbf{V}$, (i.e. the vergence error).

The *maximum residual error*, $\mathbf{E}_{R\max}$, thus obtained for all ten virtual world points and all combinations of parameter errors in a given experiment was used as a measure of the global surface measurement error for that experiment. i.e.:

$$\mathbf{E}_{R\max} = \max_{c=0}^{c<C}\max_{i=1}^{i\le 10}\left[(\hat{\mathbf{P}}_{i,c} - \hat{\mathbf{V}}_c) - (\mathbf{P}_i - \mathbf{V})\right] \tag{6.3}$$

It is clear that the residual error at a point, as defined in equation 6.2, will be zero at the vergence point and tend to increase with its distance from the vergence point. Thus, the ten points at the extremes of the viewing volume that were chosen for investigation were those most likely to exhibit the maximum possible residual error.

6.3. Individual parameter analysis

6.3.1. Aim

The aim of this experiment was to characterise individually the surface measurement error induced by the introduction of each of the external parameters into the ASP stereo system model. The results of this analysis would form the basis of the specification of the final ASP sensor head. The capability of this specification for the sensor head to meet the ASP stereo vision system 3D measurement accuracy requirements would then be tested in the next experiment, described in section 6.4.

6.3.2. Method

The ASP stereo system model was initially set up in the basic configuration that was established in the previous chapter, i.e. with both camera focal lengths set to 50mm and a camera baseline of 300mm. This camera baseline was provided by centring the Camera One and Camera Two azimuth actuators at world co-ordinates (−150mm,0,0) and (+150mm,0,0) respectively. There was no translation between the azimuth actuators, elevation actuators or cameras and the only rotations present were those necessary to achieve vergence.

The basic analytic procedure that was described in the previous section was then carried out with the cameras verged to the world point (0,0,5m). The experiment was carried out for seven different combinations of external parameters:

1. azimuth actuator rotation: $\hat{\Phi}_{A1} = \Phi_{A1} \pm \varepsilon_{\Phi_A}$, $\hat{\Phi}_{A2} = \Phi_{A2} \pm \varepsilon_{\Phi_A}$;
2. elevation actuator rotation: $\hat{\Omega}_{E1} = \Omega_{E1} \pm \varepsilon_{\Omega_E}$, $\hat{\Omega}_{E2} = \Omega_{E2} \pm \varepsilon_{\Omega_E}$;
3. *X*-component of azimuth actuator translation: ${}^{W}\hat{X}_{A1} = {}^{W}X_{A1} \pm \varepsilon_{w_{X_A}}$, ${}^{W}\hat{X}_{A2} = {}^{W}X_{A2} \pm \varepsilon_{w_{X_A}}$;
4. *Y*-component of azimuth actuator translation: ${}^{W}\hat{Y}_{A1} = {}^{W}Y_{A1} \pm \varepsilon_{w_{Y_A}}$, ${}^{W}\hat{Y}_{A2} = {}^{W}Y_{A2} \pm \varepsilon_{w_{Y_A}}$;
5. *Z*-component of azimuth actuator translation: ${}^{W}\hat{Z}_{A1} = {}^{W}Z_{A1} \pm \varepsilon_{w_{Z_A}}$, ${}^{W}\hat{Z}_{A2} = {}^{W}Z_{A2} \pm \varepsilon_{w_{Z_A}}$;

6. azimuth actuator off-axis wobble: ω_{A1}, $\phi_{A1} = -180°,-165°,\cdots,165°,180°$, ω_{A2}, $\phi_{A2} = -180°,-165°,\cdots,165°,180°$;

7. elevation actuator off-axis wobble: κ_{E1}, $\omega_{E1} = -180°,-165°,\cdots,165°,180°$, κ_{E2}, $\omega_{E2} = -180°,-165°,\cdots,165°,180°$.

As the ASP sensor head model was to be set up with only the small rotations required to achieve vergence and with no additional translation between the azimuth actuators, elevation actuators or cameras. Therefore the effect of many of the parameter would be very similar and it was deemed unnecessary to perform the analysis for every one of the errors listed in sub-section 6.1.3. The results obtained for the seven sets of parameters that were investigated would be broadly representative of the complete list of parameters.

The experiment would, ideally, have been repeated for all of the vergence points investigated in the previous chapter. However this was precluded by the length of time required for the analysis of each parameter. Therefore only the vergence point (0,0,5m) was considered.

6.3.3. Results

The results of the experiments described above are given in the graphs plotted in figure 6.6 to figure 6.17. These graphs show maximum relative world error, $\mathbf{E}_{R\max}$, and maximum vergence error, $\mathbf{E}_{V\max}$, respectively plotted against: azimuth actuator rotation error, ε_{Φ_A} (figure 6.6 and figure 6.7); elevation actuator rotation error, ε_{Ω_E} (figure 6.8 and figure 6.9); azimuth actuator X translation error, $\varepsilon_{w_{X_A}}$ (figure 6.10 and figure 6.11); azimuth actuator Z translation error, $\varepsilon_{w_{Z_A}}$ (figure 6.12 and figure 6.13); azimuth actuator "off-axis wobble", ω_A (figure 6.14 and figure 6.15); elevation actuator "off-axis wobble", κ_E (figure 6.16 and figure 6.17). No graphs have been presented for azimuth actuator Y translation error, $\varepsilon_{w_{Y_A}}$, because its effect on maximum relative world error was negligibly small in comparison to the X and Z components.

It should also be noted that a different scaling has been used to plot the graphs shown in figure 6.14 to figure 6.17. This was necessary in order to adequately accommodate the step size of 0.01° that was used in the simulation of "off-axis wobble".

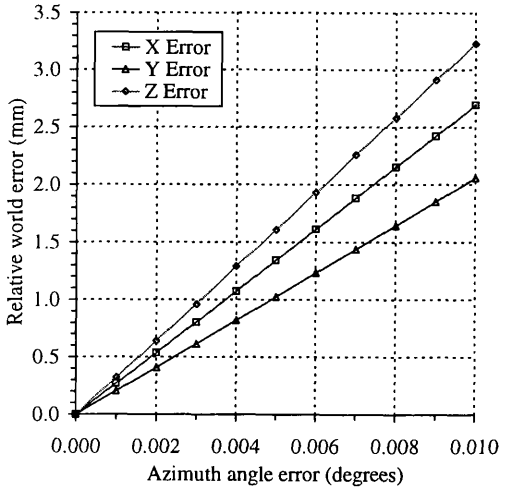


Figure 6.6. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. azimuth actuator rotation error, ε_{Φ_A}.

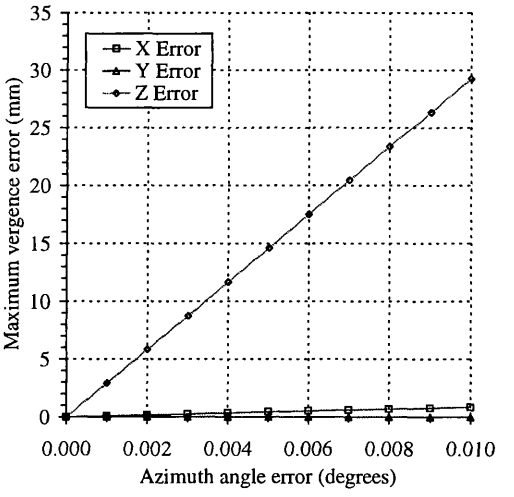


Figure 6.7. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. azimuth actuator rotation error, ε_{Φ_A}.

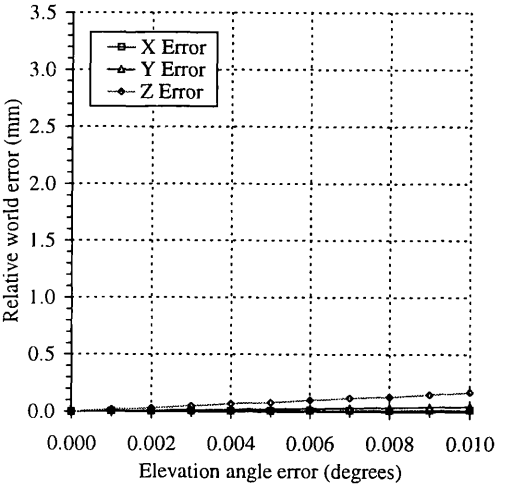


Figure 6.8. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. elevation actuator rotation error, ε_{Ω_E}.

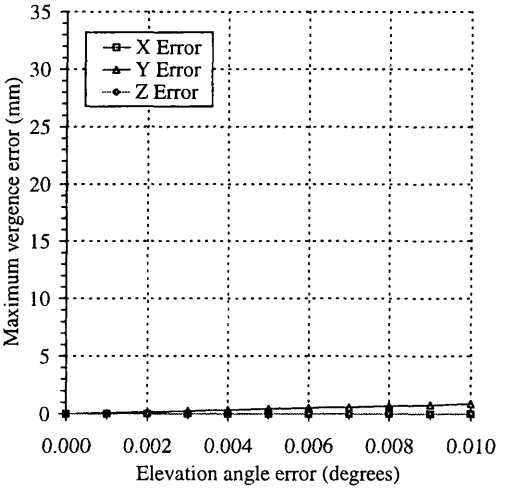


Figure 6.9. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. elevation actuator rotation error, ε_{Φ_A}.

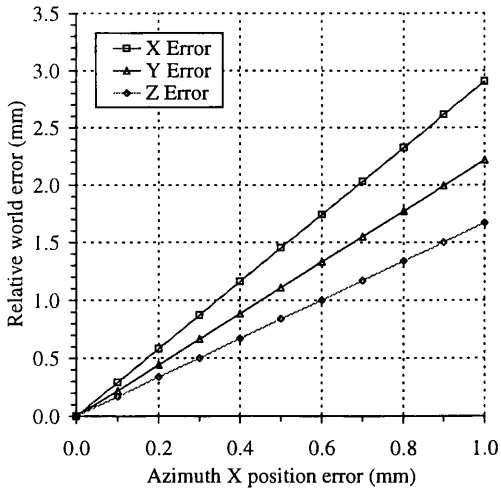


Figure 6.10. Graph of maximum residual error, $\mathbf{E}_{R\max}$*, vs. azimuth actuator* X *translation error,* $\varepsilon_{w_{X_A}}$.

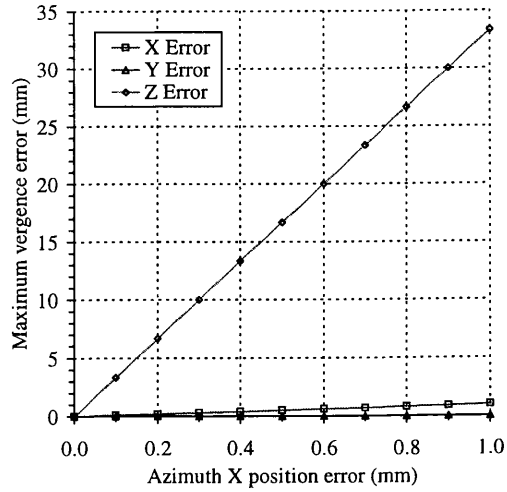


Figure 6.11. Graph of maximum vergence error, $\mathbf{E}_{V\max}$*, vs. azimuth actuator* X *translation error,* ε_{Φ_A}.

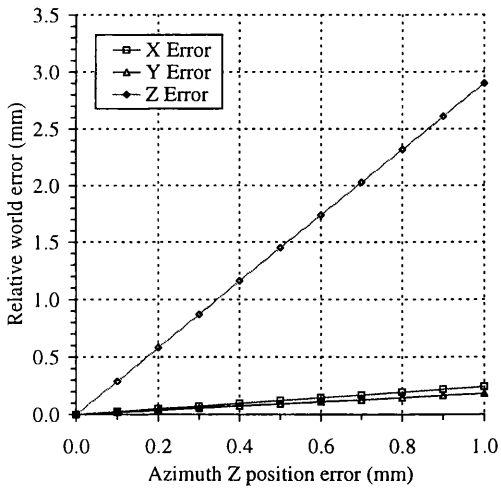


Figure 6.12. Graph of maximum residual error, $\mathbf{E}_{R\max}$*, vs. azimuth actuator* Z *translation error,* $\varepsilon_{w_{Z_A}}$.

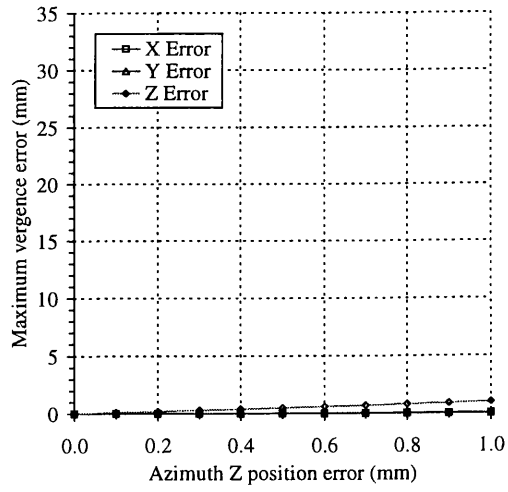


Figure 6.13. Graph of maximum vergence error, $\mathbf{E}_{V\max}$*, vs. azimuth actuator* Z *translation error,* $\varepsilon_{w_{Z_A}}$.

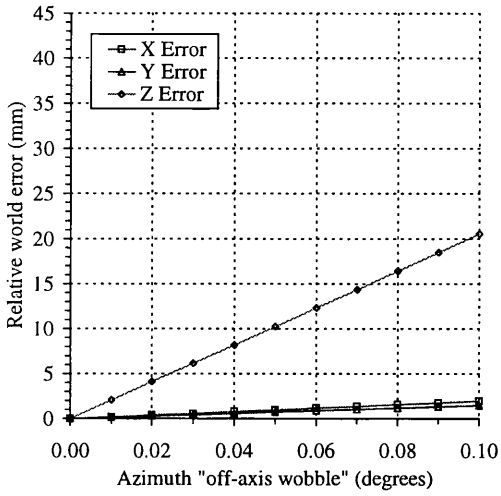


Figure 6.14. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. azimuth actuator "off-axis wobble", ω_A.

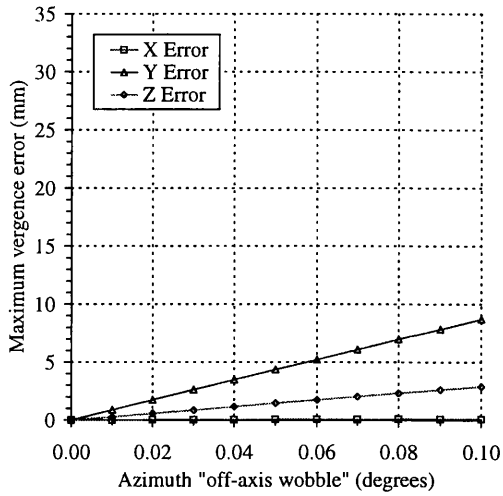


Figure 6.15. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. azimuth actuator "off-axis wobble", ω_A.

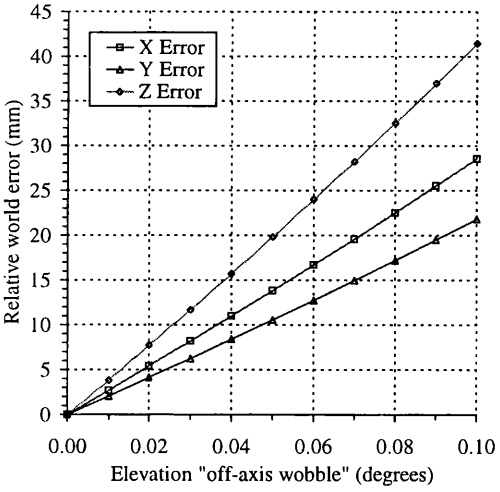


Figure 6.16. Graph of maximum residual error, $\mathbf{E}_{R\max}$, vs. elevation actuator "off-axis wobble", κ_E.

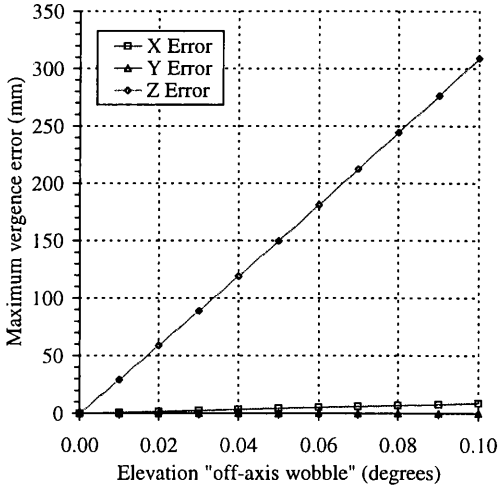


Figure 6.17. Graph of maximum vergence error, $\mathbf{E}_{V\max}$, vs. elevation actuator "off-axis wobble", κ_E.

6.3.4. Interpretation of results

The objective of this experiment was to characterise the individual influence of the major sources of sensor position error in the ASP sensor head on 3D measurement accuracy. This analysis would form the basis of the specification of the final ASP sensor head so that it would be capable of meeting the specified 3D measurement accuracy requirements.

The minimum acceptable 3D measurement accuracy for the ASP stereo vision system was specified, in the performance requirements set out in Chapter 3, to be ±5.0mm. This figure represents the maximum acceptable relative world error induced with all error parameters acting together. If this target is to be achieved with all the error parameters acting simultaneously, then it is clear that a considerably smaller relative world error must be induced when each error parameter acts in isolation. It was estimated that an individual maximum relative world error of about 0.25mm would be required. This figure was therefore chosen as the basis for the initial specification of components for the ASP sensor head.

It can be seen from the graphs of figure 6.6, figure 6.14 and figure 6.16 that the target for individual maximum relative world error of 0.25mm is met by an azimuth rotation error, ε_{Φ_A}, azimuth actuator "off-axis wobble", ω_A, and elevation actuator "off-axis wobble", κ_E, of about 0.001°. In isolation, an elevation rotation error, ε_{Ω_E}, of 0.001° was found to induce considerable less maximum relative world error than the target figure of 0.25mm. The graphs of figure 6.10 and figure 6.12 show that the target maximum relative world error is achieved by X and Z translation errors of about 0.1mm. The Y-component of translation error was found to have much less effect on the maximum relative world error than the X- and Z-components.

It is interesting to note that the all of the graphs in figure 6.6 to figure 6.17 appear linear. This suggests that an alternative method for performing the experiments in this chapter might have been to use some form of perturbation or linearised error analysis.

6.4. Combined parameter analysis

6.4.1. Aim

The preceding experiment established that, in isolation, a value of 0.001° for the rotational error parameters of the ASP sensor system model induced a maximum relative world error of no more than about 0.25mm. Similarly, a value of 0.1mm for the translational error parameters also induced a maximum relative world error of no

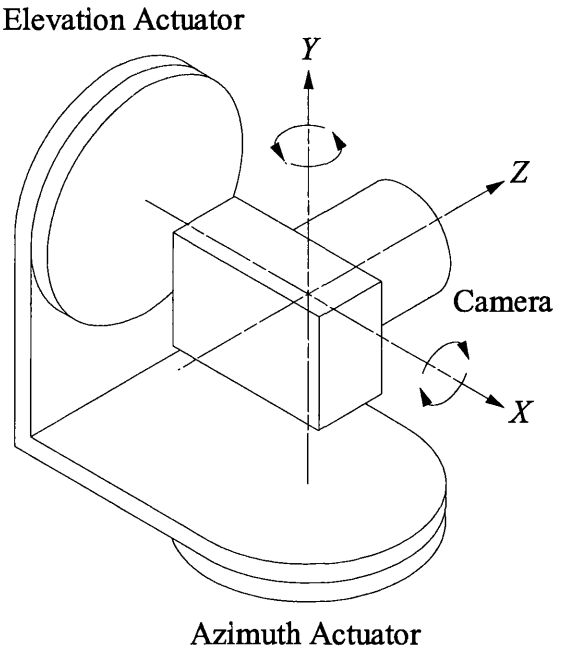


Figure 6.18. An actuated camera.

more than 0.25mm, when these error parameters act in isolation. The aim of this experiment was to establish the maximum relative world error induced when the errors in all of the sensor position parameters act simultaneously. A suitable configuration for the ASP sensor could then be derived by iteratively adjusting appropriate parameter errors as necessary until the specified maximum 3D measurement error of 5.0mm is obtained.

6.4.2. Actuator configuration

All of the previous simulation experiments have assumed an idealised sensor head configuration. The only translations that have been introduced into the ASP sensor head model have been those required to provide the stereo camera baseline, i.e. ${}^{W}X_{A1} = -150\text{mm}$ and ${}^{W}X_{A2} = +150\text{mm}$, with all other translations zero (except for the small translation error when relevant). However figure 6.18 illustrates that the mechanical constraints inherent in an actual physical sensor head require additional translations. In particular, translations are required between the azimuth and elevation actuators and between the elevation actuator and camera to allow them to be mounted together.

A series of auxiliary simulation experiments were carried out to establish if these extra translations, when combined with rotational errors, introduced a significant amount of additional measurement error. It was found that the additional error

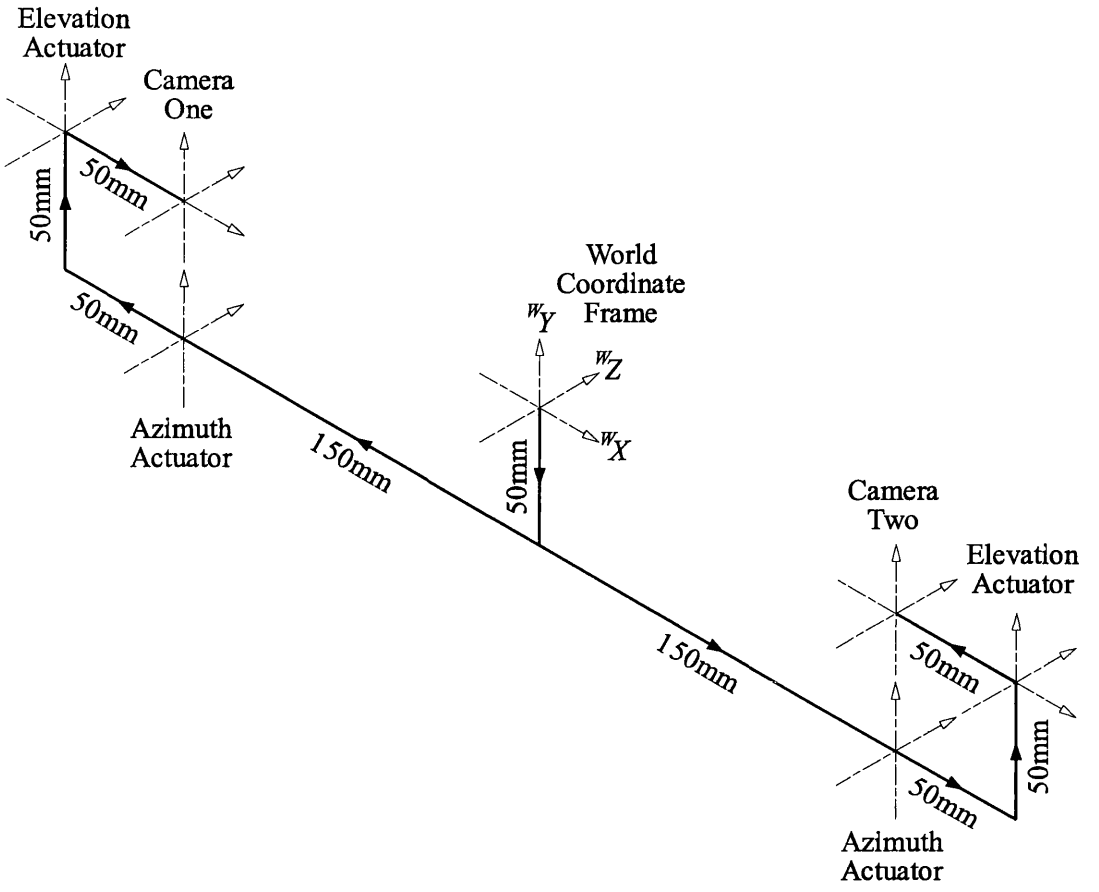


Figure 6.19. Schematic of the translations introduced into the ASP sensor head model.

introduced by translations of up to 100mm and rotational errors of 0.001° was negligibly small. However, for completeness, the additional translations that are illustrated in figure 6.19 were added into the ASP sensor system model anyway. A distance of 50mm was chosen for these additional translations since this was thought to be approximately the distance that would be used in the final ASP sensor head.

6.4.3. Method

The objective of this experiment was to establish the maximum relative world error that would be induced when all of the error parameters act simultaneously. Table 6.2 summarises the "actual" value assigned to the parameters in the ASP sensor system model and the range of error values associated with each. In the table, Φ_{V1}, Φ_{V2}, Ω_{V1}, and Ω_{V2} are the azimuth and elevation rotations required to bring Camera One and Camera Two to convergence. The other parameters are as previously defined.

Parameter	Actuator/Camera		
	Azimuth Actuator	Elevation Actuator	Camera
Rotation	$\hat{\Phi}_{A1} = \Phi_{V1} \pm 0.001°$ $\hat{\Phi}_{A2} = \Phi_{V2} \pm 0.001°$	$\hat{\Omega}_{E1} = \Omega_{V1} \pm 0.001°$ $\hat{\Omega}_{E2} = \Omega_{E2} \pm 0.001°$	$\hat{\mathrm{K}}_{C1} = \pm 0.001°$ $\hat{\mathrm{K}}_{C2} = \pm 0.001°$
"Wobble" Direction	$\phi_{A1} = [-180°, +180°]$ $\phi_{A2} = [-180°, +180°]$	$\omega_{E1} = [-180°, +180°]$ $\omega_{E2} = [-180°, +180°]$	$\kappa_{C1} = [-180°, +180°]$ $\kappa_{C2} = [-180°, +180°]$
"Wobble" Angle	$\omega_{A1} = 0.001°$ $\omega_{A2} = 0.001°$	$\kappa_{E1} = 0.001°$ $\kappa_{E2} = 0.001°$	$\phi_{C1} = 0.001°$ $\phi_{C2} = 0.001°$
X-Translation	${}^{W}\hat{X}_{A1} = -150 \pm 0.1\text{mm}$ ${}^{W}\hat{X}_{A2} = +150 \pm 0.1\text{mm}$	${}^{A1}\hat{X}_{E1} = -50 \pm 0.1\text{mm}$ ${}^{A2}\hat{X}_{E2} = +50 \pm 0.1\text{mm}$	${}^{E1}\hat{X}_{C1} = +50 \pm 0.1\text{mm}$ ${}^{E2}\hat{X}_{C2} = -50 \pm 0.1\text{mm}$
Y-Translation	${}^{W}\hat{Y}_{A1} = -50 \pm 0.1\text{mm}$ ${}^{W}\hat{Y}_{A2} = -50 \pm 0.1\text{mm}$	${}^{A1}\hat{Y}_{E1} = +50 \pm 0.1\text{mm}$ ${}^{A2}\hat{Y}_{E2} = +50 \pm 0.1\text{mm}$	${}^{E1}\hat{Y}_{C1} = \pm 0.1\text{mm}$ ${}^{E2}\hat{Y}_{C2} = \pm 0.1\text{mm}$
Z-Translation	${}^{W}\hat{Z}_{A1} = \pm 0.1\text{mm}$ ${}^{W}\hat{Z}_{A2} = \pm 0.1\text{mm}$	${}^{A1}\hat{Z}_{E1} = \pm 0.1\text{mm}$ ${}^{A2}\hat{Z}_{E2} = \pm 0.1\text{mm}$	${}^{E1}\hat{Z}_{C1} = \pm 0.1\text{mm}$ ${}^{E2}\hat{Z}_{C2} = \pm 0.1\text{mm}$

Table 6.2. "Actual" parameter values and errors used in the combined parameter analysis experiments.

Table 6.2 illustrates the complexity of the analysis to be undertaken with a total of thirty parameters to be varied during the analysis. Of these thirty, the 24 parameters for actuator/camera rotation and translation were each to take either positive or negative error values and the six "off-axis wobble" direction parameters were each to cover a full 360° in 36 × 10° steps. The total number of permutations of the model for all of these parameter errors is therefore (overleaf):

$$\text{number of permutations} = 2^{24} \times 36^{6} \approx 4 \times 10^{16} \tag{6.4}$$

Each iteration of the simulation to calculate relative world error took approximately one second to compute. An exhaustive search for the maximum residual error at a single world point for all permutations of parameter errors would consequently have taken approximately one billion years! Therefore, the exhaustive search approach that had been used in the analysis of individual parameter errors was clearly impractical for use in the combined parameter analysis experiments.

An alternative approach for the introduction of the parameter errors into the model was therefore devised. Instead of exhaustively searching through all of the permutations of parameter errors, the errors were introduced into the model in small sub-sets. The combination of error values that induced the maximum relative world error for a particular sub-set was established and the error parameters of that sub-set were then fixed at these values for the remainder of the analysis. The next sub-set of parameter errors, following the mechanical linkages of the model, was then introduced into the simulation. The combination of parameter errors in the new sub-set that induced the maximum cumulative relative world error was then analysed and so on until all the parameter errors had been introduced into the model.

The first sub-set of parameter errors to be introduced into the simulation were the errors of ±0.1mm in the *X*- *Y*- and *Z*-components of azimuth actuator translation for both Camera One and Camera Two. All 64 permutations (2^6) of error for these six parameters (three components for Camera One and three for Camera Two), with all other parameter errors set to zero, were searched to find the combination that induced the maximum *Z*-component of relative world error. This combination of errors was assigned to the azimuth actuator translations for the remainder of the analysis.

The errors of ±0.001° in the azimuth actuator rotation angles for both actuators comprised the second sub-set of parameter errors to be introduced into the model. The four possible combinations of error were investigated while the azimuth actuator translation errors were set to the previously determined values and all the other parameter errors were set to zero. Thus the specific combination of azimuth actuator rotation errors and azimuth translation errors that induced the maximum Z-component of relative world error was found. The process was then repeated for the third sub-set of parameter errors which comprised the 0.001° "off-axis wobble" for both azimuth actuators. The "wobble direction" for each camera was introduced in 36 steps of 10°; there was therefore a total of 1296 (36^2) error permutations for the third sub-set of errors.

The fourth, fifth and sixth sub-sets of errors to be introduced into the model comprised, respectively: ±0.1mm in the *X*- *Y*- and *Z*-components of translation for both elevation actuators; ±0.001° in both elevation actuator rotation angles; and 0.001° of "off-axis wobble" for both elevation actuators, again introduced in 10° steps of "wobble direction". The final three sub-sets of parameter errors followed the same pattern except they were introduced into both sets of camera parameters instead of the elevation actuator parameters.

6.4.4. Results

The absolute maximum *X*- *Y*- and *Z*-components of relative world error were obtained for all ten world points ($\mathbf{P}_1,\mathbf{P}_2,\cdots,\mathbf{P}_{10}$) and all nine of the vergence points. These results are summarised in table 6.3 overleaf.

Vergence Point	*X*-Error (mm)	*Y*-Error (mm)	*Z*-Error (mm)
(0.0, 0.0, 4.5)	1.580	1.177	2.572
(0.0, 0.0, 5.0)	1.820	1.360	2.843
(0.0, 0.0, 5.5)	2.076	1.555	3.125
(0.0, 0.5, 4.5)	1.638	1.527	2.717
(0.0, 0.5, 5.5)	2.139	1.903	3.281
(0.5, 0.0, 4.5)	2.004	1.248	2.850
(0.5, 0.0, 5.5)	2.490	1.626	3.421
(0.5, 0.5, 4.5)	2.083	1.653	3.007
(0.5, 0.5, 5.5)	2.572	2.022	3.587

Table 6.3. Absolute maximum relative world error for each vergence point.

6.4.5. Interpretation of results

It can be seen from table 6.1 that the absolute maximum relative world error for all three components of error occurs for the vergence point (0.5m, 0.5m, 5.5m). The actual value of the error was found to be 2.572mm, 2.022mm and 3.587mm for the *X*-, *Y*- and *Z*-components respectively. All of these values are well within the required maximum residual error of 5.0mm. It can therefore be concluded that the specification of a minimum actuator rotation accuracy of ±0.001°, a maximum actuator "off-axis wobble" of 0.001° and a maximum translational tolerance of ±0.1mm is sufficient to provide acceptably accurate positional feedback from the ASP sensor head.

It should be noted that the expected absolute maximum residual error figures of be 2.572mm, 2.022mm and 3.587mm are indeed the maximum expected values. In order for an error of this magnitude to occur in the actual system, all of the components of the ASP sensor head would have to be at the very limits of their tolerance, exactly the right combination of errors would have to occur (with a probability of approximately one in 4×10^{16}) and the measured point would have to be at one of the extreme corners of the viewing volume. Therefore the actual RMS error that would be

introduced into surface measurements was expected to be considerably less than the predicted absolute maximum residual error.

6.5. Conclusions

6.5.1. Review of objective

The objective of the experiments described in this chapter was to analyse the effect on 3D measurement accuracy of errors in the mechanical components and construction of the proposed ASP sensor head. This would enable the mechanical tolerances required to meet the specified accuracy requirements to be determined. A detailed mechanical design for the ASP sensor head that meets the resulting tolerance requirements, assuming they were not prohibitively high, could then be drawn up. This would allow the ASP sensor head to be fabricated with a high degree of confidence that it would provide sufficiently accurate sensor position and orientation information to meet the surface measurement accuracy requirements.

6.5.2. Summary of results

Prior to the experiments, three significant types of mechanical errors were identified in the proposed design of the ASP sensor head: the (in)accuracy of the positional feedback from the rotational actuators, "off-axis wobble" in the rotational actuators and the (mainly translational) tolerances inevitable in machining and assembling mechanical components. The first set of simulation experiments, described in section 6.3, applied each type of error to the ASP sensor head model in isolation. It was thus determined that translational tolerances of approximately ±0.1mm and rotational tolerances of approximately ±0.001° induced a significant amount of 3D surface measurement error.

The second set of simulation experiments, described in section 6.4, analysed the effect on measurement accuracy of applying all of the above tolerances to the ASP sensor model simultaneously (as they would act in the actual system). It was found that the resulting maximum relative measurement error for the full extent of the ASP working volume was 3.587mm. This figure is within the minimum accuracy of ±5.0mm acceptable for the ASP stereo vision system. The specification of maximum translational tolerances of ±0.1mm and rotational tolerances of ±0.001° (±3.6arcsec) was therefore adopted for the selection of components and the construction of the ASP sensor head. Details of the design and fabrication of the final ASP sensor head based upon these tolerances are given in Chapter 7.

6.5.3. Conclusions

The approach to dynamic stereo vision that has been proposed in this work relies upon accurate feedback of the camera positions and orientation to facilitate accurate surface measurement. Prior to these experiments, there was no detailed data available regarding the levels of feedback accuracy and mechanical tolerances required to achieve a given level of surface measurement accuracy. The design of a sensor head capable of providing a known level of performance has therefore been impossible and many researchers have dismissed this approach [Thacker, 1992]. Those few who have previously attempted to use this direct feedback approach have relied upon little more than educated guesswork and heuristics to design their sensor systems. This has resulted in completed systems that are suitable, at best, for only very low accuracy measurement [Krotkov, 1989].

The simulation experiments described in this chapter, have introduced a new rigorous design methodology that is intended to re-establish the feasibility of the direct feedback approach. These experiments have allowed the design of a dynamic stereo sensor head with specified performance limits. The result has been a full specification for the mechanical components and construction tolerances required to permit a given level of 3D measurement accuracy. This has in turn led to the fabrication of a high precision, well characterised, stereo sensor head as detailed in Chapter 7.

Of course, the other components of the stereo vision system (e.g. the stereo matcher and camera calibration) also have a minimum level of performance to permit a given level of 3D measurement accuracy. Indeed it would have been beneficial to include in the simulations the effects of factors such as the stereo matching accuracy or the amount of uncorrected lens distortion. Hence a specification for these requirements could also have been established. Unfortunately, time was not available to extend the ASP sensor system model to accommodate these parameters. Even so, the experiments described in this chapter have ensured that, in theory at least, the mechanical aspects of the completed ASP sensor head are of a high enough specification to allow the desired level of 3D measurement accuracy. Analysis of the 3D measurement accuracy actually achieved in practice is the subject of later chapters.

It is worthy of note that the ASP sensor model played a critical part in the experiments described in this chapter. The in-depth analysis that was required simply could not have been performed without a modelling tool similar to the ASP sensor model.

Chapter 7
Sensor system fabrication

7.1. Introduction

The objective of the work described in this chapter was the fabrication of the ASP sensor system. Chapter 4 of thesis presented a model for the proposed ASP sensor head. This model was then used in a series of simulation experiments that are described in Chapters 5 and 6. The result of these simulation experiments was a specification for the fabrication of the actual ASP sensor head. This chapter reports how this fabrication was carried out and describes the completed ASP sensor system produced by integrating the sensor head with the frame grabber and other computer hardware required to support the vision system software.

7.2. Selection of actuators

The simulation experiments described in Chapter 6 were aimed at investigating the relationship between surface recovery accuracy and the accuracy of actuator feedback and tolerances in the construction of the ASP sensor head. The specific objective of these experiments was to determine a specification for the ASP sensor head that would meet the specified minimum surface recovery accuracy of 5.0mm at an object distance of approximately 5.0m. It was established that rotational actuators with a minimum accuracy of ±0.001° (±3.6 arcsec) and a maximum axial deviation of 0.001° (3.6 arcsec) were required to meet this specification. The first stage in the fabrication of the ASP sensor head was therefore to select components of a suitably high specification, a compact size and an affordable price.

A thorough review of the available range of off-the-shelf rotational actuators was carried out. Those manufactured by Physik Instrumente (P.I.) were found to be the only ones that could provide the required performance within the available budget. P.I. M-038.00 rotation stages were chosen for actuation of the azimuth axes and the slightly smaller M-037.00 stages were chosen for actuation of the elevation axes. Both types of rotation stage are driven by a D.C. motor through an integrated back-lash free

reduction gearbox and worm-wheel assembly (P.I. part number M-445.20 for the M-038.00 rotation stage and M-444.20 for the M-037.00). The gear box increases the resolution of the attached encoder to 0.6 arcsec, although this also reduces the maximum speed to $5.7°s^{-1}$. The maximum axial deviation of the stages specified by P.I. is 1.5 arcsec. These actuators therefore met the axial deviation requirements and it was expected that their high resolution and back-lash free construction would provide a positioning accuracy within the required ±3.6 arcsec.

An additional factor favouring the selected rotation stages was that their D.C. motors can be driven directly from a relatively inexpensive expansion card (P.I. part number C-812.00) inserted into a standard P.C. Comparable stepper motor systems not only fail to provide sufficient resolution, but also require special external power supplies that add greatly to the expense of the complete system.

7.3. Sensor head assembly

Sony XC77RR-CE cameras were selected for use in the Active Stereo Probe as described in Chapter 5. Having selected these cameras and the rotational actuators described above, the final stage in the fabrication of the ASP sensor head was to assemble the components. The main considerations here were to provide the required stereo baseline of 300mm and to ensure that the whole system was assembled within the translational tolerance of ±0.1mm that had been established as being significant in Chapter 6.

A length of high rigidity optical bench was chosen to provide the base-plane for the entire sensor head assembly to ensure the minimum amount of bending and hence deviation between the axes of the two azimuth actuators. P.I. was again chosen as the manufacturer of this component (P.I. part number X-050.20) because this allowed the azimuth rotation stages to be mounted directly onto the optical bench via standard carriers (B-012.20). Great flexibility was thus also obtained since a range of baselines, including the required 300mm, could be achieved by simply sliding the carriers along the optical bench.

Manufacture of the brackets required to attach and correctly orientate the azimuth and elevation actuators and the elevation actuators and cameras was contracted out to a tool machining company. The completed brackets were designed to ensure high rigidity and stability and were precisely machined to an accuracy of 0.002" (approximately 50μm). It was therefore expected that the complete ASP sensor head would be assembled well within the required translational tolerance of ±0.1mm. Figure 7.1 on page 145 shows one of the assembled actuated camera sub-assemblies.

7.4. Projection system integration

An additional consideration in the design of the ASP sensor head was the integration of the novel LCD projection system that was being developed as a separate part of the ASP project [McDonald, 1994]. The main requirement for the integration of this system into the ASP sensor head was that the illumination pattern projected by the LCD system should be steerable to track, and maximally overlap with, the fields of view of the cameras. The most effective solution found to this problem was to project the illumination from the LCD system onto a steerable mirror situated mid-way between the stereo cameras.

For simplicity, it was decided to use the same combination of P.I. rotation stages for the actuation of the mirror as used for the cameras, the only difference being that a modified design for the brackets was required to accommodate the slightly greater size of the mirror. Again it was straightforward to mount the whole actuated mirror sub-assembly onto a P.I. carrier and thence onto the optical bench between the cameras. Since the P.I. C-812.00 D.C. motor controller is only capable of driving four axes, an additional controller had to be purchased to drive the mirror actuators. Figure 7.2 shows the completed ASP sensor head including the actuated projection mirror, while figure 7.3 shows the LCD projection system.

Although the LCD projection system was originally intended to provide the textured scene illumination for stereo capture, it was later found that insufficient illuminating power was provided by the device. A custom-built flash illumination system was then tested as an alternative source of textured illumination. However this was also found to provide insufficient illumination power, with the texture pattern appearing only very faintly on the surface of illuminated objects. Details of the experiments carried out using the LCD and flash projection systems can be found in [McDonald, 1994].

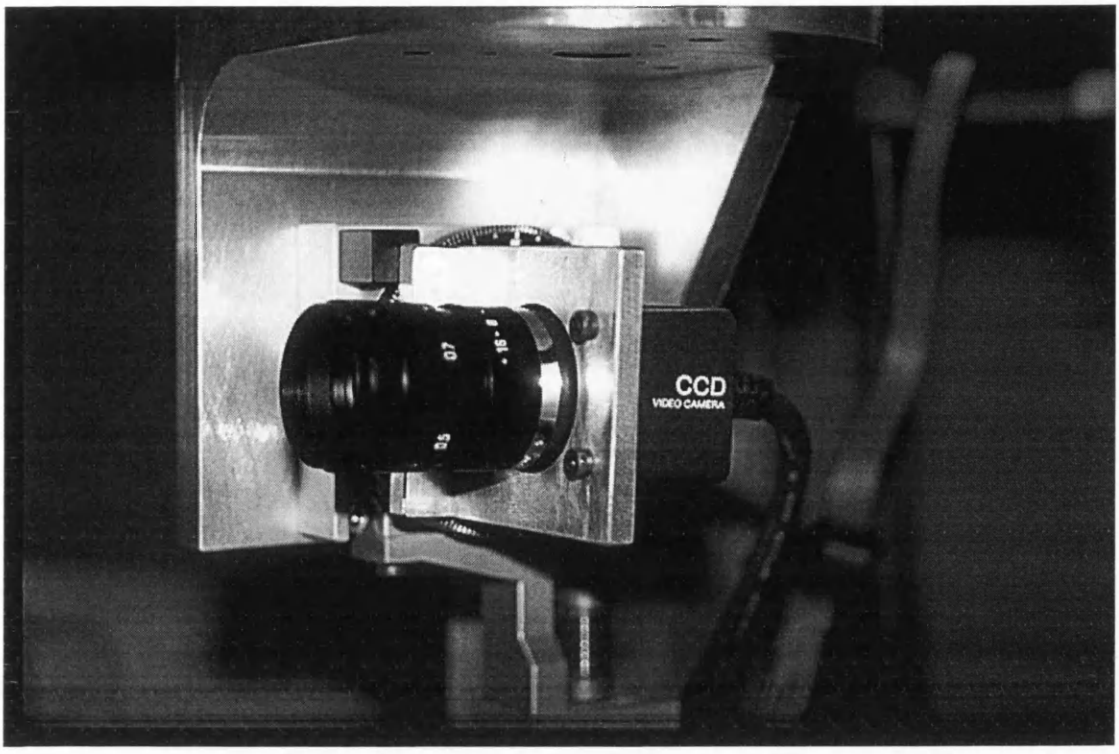


Figure 7.1. Close-up of one of the actuated cameras.

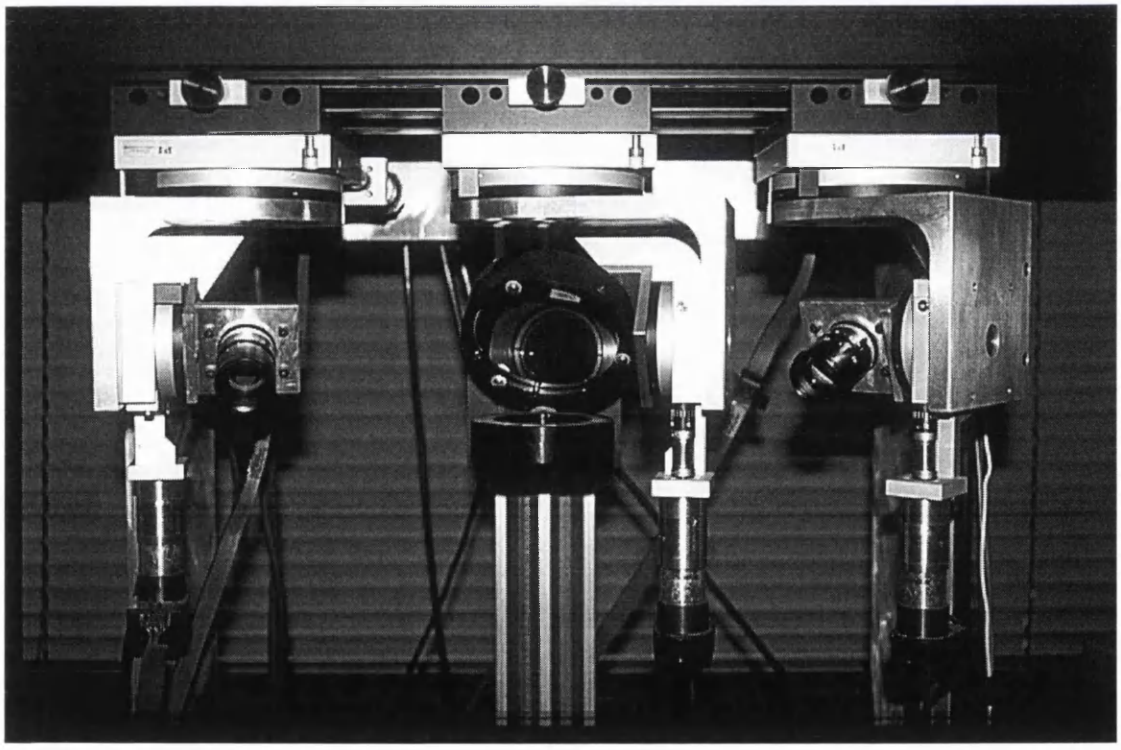


Figure 7.2. The completed ASP sensor head including the actuated projection mirror.

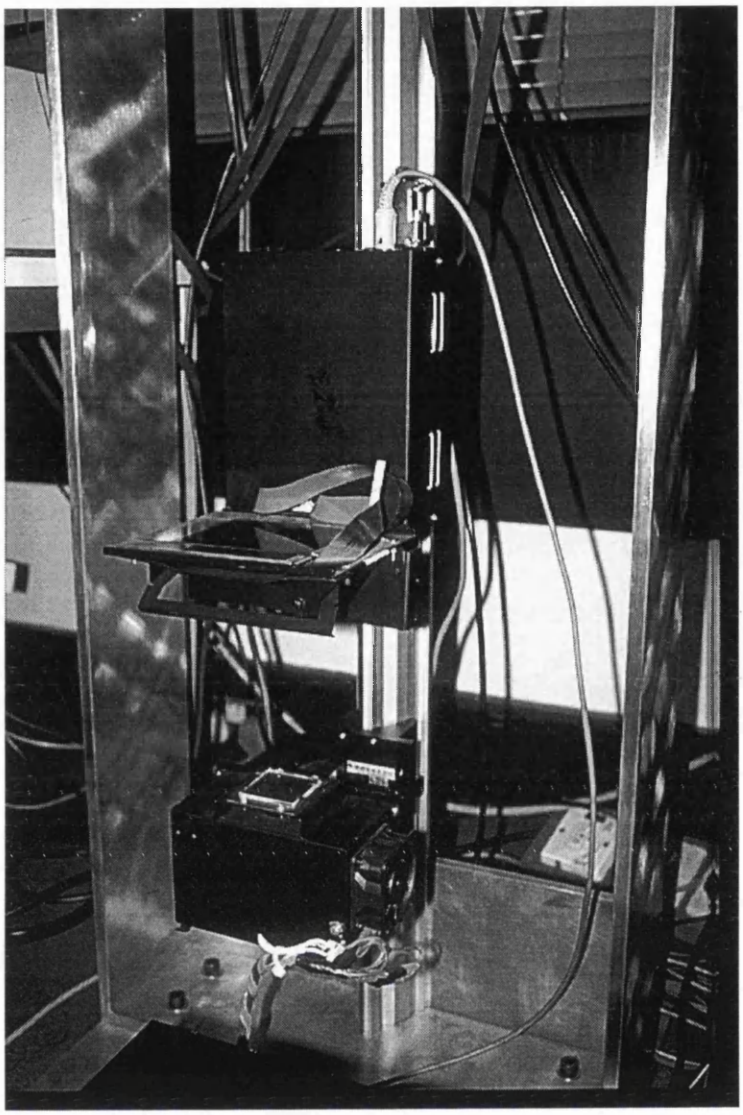


Figure 7.3. The LCD projection system.

Figure 7.4. The 35mm slide projector used in the ASP sensor system.

Figure 7.5. The complete ASP sensor system.

Eventually a standard 35mm slide projector was introduced into the system as this was known to provide adequate illuminating power. However the gravity fed slide change mechanism meant that the projector could not been mounted vertically as was required to reflect its beam off the projection mirror. Therefore the ability to steer the illumination beam that was possible with the LCD projector and flash illumination devices was lost. However this problem was circumvented by using a projection lens with a fixed field of view larger than that of the cameras. Hence the narrower fields of view of the cameras could be re-directed while still remaining within the extended field of view of the projector. The 35mm slide projector used is shown in figure 7.4.

7.5. Full system integration

Figure 7.5 shows the completed ASP sensor system and a schematic of the system is shown in figure 7.6. The main processing unit for the system is the host workstation. Originally this was a two processor Sun SPARCstation 10, but due to a combination of unfortunate circumstances this had to be replaced by a less powerful Sun SPARCstation LX. The stereo image capturing facility required by the ASP system is provided by a DataCell S2200 frame-store housed within the host workstation. A DataCell S2200 [DataCell, 1992a; DataCell 1992b] frame-store was chosen primarily because it offered separate red, green and blue frame buffers each of full CCIR resolution and because it was compatible with the S-bus of modern Sun workstations.

The left camera of the sensor head is connected to the green video input of the frame-store and the right camera is connected to the red video input. The left camera provides the sync. signal for the right camera and the frame-store. The green and red output channels of the frame-store are connected to an RGB (red/green/blue) monitor to provide live stereo anaglyph output of the camera images. Appropriate software drivers were developed to allow the frame-store to freeze, capture and display images from either or both of the cameras live on the RGB monitor.

The P.C. is included in the ASP system to house the two P.I. actuator controller cards and a custom designed projector controller card [McDonald, 1994]. Low-level software drivers were developed to control all six of the actuated axes of the sensor head. A communications protocol was also developed to allow control messages to be passed between the host workstation and the P.C. via an RS232 interface.

The development of the software components for the 3D computer vision system is the subject of Chapters 8, 9 and 10 of this thesis. A fairly rudimentary user interface was developed initially to allow the ASP sensor system to be controlled from the host workstation. This user interface was later superseded by a more sophisticated one

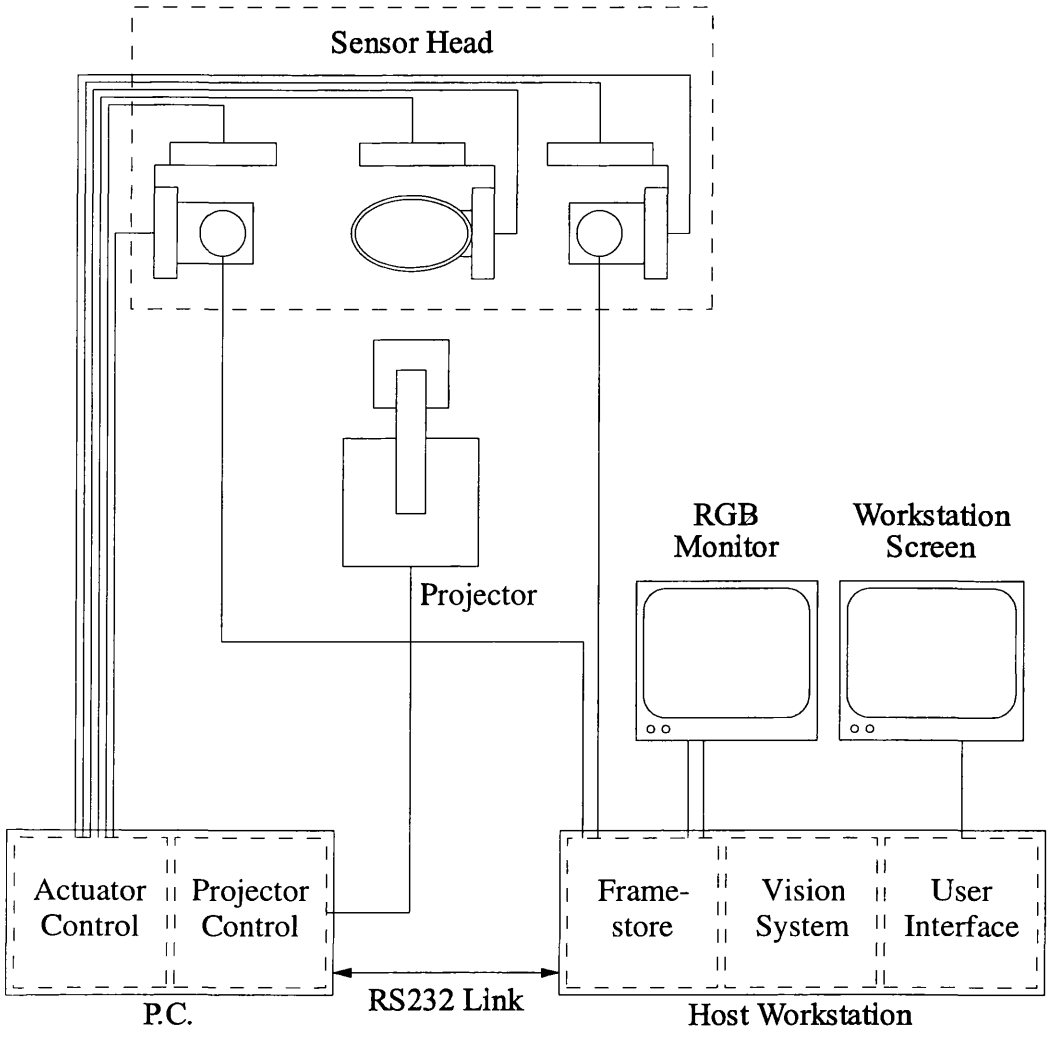


Figure 7.6. Schematic of the complete ASP sensor system.

developed for the CopyCAD demonstration system. Chapter 11 describes the development of this graphical user interface and its integration with the vision system software and ASP sensor system to produce the final prototype of the CopyCAD demonstrator.

7.6. Performance verification

Having completed the fabrication of the ASP sensor head, it was important to verify that its actual performance matched that predicted by the ASP sensor system model.

In order to do this accurately, it would be necessary to set-up the ASP sensor system model with an accurate representation of the real ASP sensor head. This would require the sensor head to be fully calibrated, and was therefore not yet possible. However it

was possible to verify that the viewing volume and disparity range of the actual sensor head was within the specified performance requirements.

7.7. Conclusions

It can be seen from the discussion presented in this chapter that the adopted approach of extensive characterisation of the ASP sensor head using computer simulations prior to its actual fabrication has greatly simplified the final design process. Using the specification that was determined from the simulation experiments it was a relatively simple process to select suitable components and have them assembled to the required tolerances. Having selected components that are expected to surpass the specified requirements, a high level of confidence can then be placed in the completed system meeting or even surpassing the required level of sensor position feedback accuracy.

The specified level of sensor position feedback accuracy was determined to ensure that the completed system would meet the required minimum 3D measurement accuracy (±5.0mm at an object distance of approximately 5.0m) when operating dynamically in open-loop mode. In open-loop mode of operation it is proposed that the external orientation parameters of the stereo system are estimated purely from the position feedback from the camera actuators. The results of the simulation experiments predict that the completed sensor system should achieve the required minimum level of 3D measurement accuracy. However, it is clear that the accuracy actually achieved in practice can only be tested after the development of the open-loop dynamic calibration scheme. Accordingly, the problem of open-loop dynamic calibration is not considered until the further work section of Chapter 12.

Perhaps a more immediate and practical objective to the development of the ASP sensor system was that it should provide a convenient and flexible system for image capture and scene illumination as the fundamental basis for the ASP project. That this was successfully achieved is demonstrated by the results of 3D surface recovery obtained from images captured using the system that are presented in Chapter 11. However, not only was the ASP sensor system used by the author as part of the CopyCAD videometrics system, but it was also used extensively by other members of the ASP project team. The most notable example is McDonald's experiments into active vision [McDonald, 1994]. In particular, his implementations of stereo scene coding and dynamic video feedback relied heavily on the ASP sensor system.

Chapter 8
Stereo matcher development

8.1. Introduction

8.1.1. Objective

A fundamental component of any stereo vision system is the stereo matcher: a computer algorithm that automatically determines correspondences between a stereo pair of images. The Multiple Scale Signal Matcher (MSSM) [Zhengping, 1988] was chosen as the stereo matcher to be used in the ASP stereo vision system for two basic reasons:

1. It represented the state-of-the-art in correlation-based stereo matchers and its performance was comparable to that of any other algorithm reported in the literature.

2. It had been developed at the Turing Institute. The original source code was therefore readily available and the algorithm's designers were easily accessible to provide advice and information about undocumented features. This made it much more feasible to consider making any required developments to the algorithm or its software implementation.

Although the original MSSM stereo matcher did represent the state-of-the-art in correlation-based stereo matchers, some aspects of its performance still fell short of the requirements for the ASP stereo vision system while others were almost completely unknown. The objective of the work described in this chapter was therefore to identify and address the deficiencies in the original version of the MSSM stereo matcher. The objective of the work described in the following chapter (Chapter 9) was then to analyse the performance envelope of the MSSM stereo matcher and to investigate its implications in the context of the ASP stereo vision system.

8.1.2. Stereo matcher requirements

It has already been mentioned in Chapter 3 that there were three basic requirements for the stereo matcher to be used within the ASP stereo vision system:

1. *Accuracy.* The minimum acceptable depth measurement accuracy for the ASP stereo vision system was given in Chapter 3 as ±5mm over the specified viewing volume depth of 0.5m and distance of between 4.5m and 5.5m. It has already been established in previous chapters that the finite accuracy of the stereo matcher will be only one of many sources of measurement error in the system. The stereo matcher must therefore be sufficiently accurate to provide depth measurements better than ±5mm.

2. *Speed.* Although real-time performance (i.e. with rates of around 25Hz) is the ultimate goal for dynamic vision systems, it was accepted that this level of performance was well beyond the limitations of the current generation of stereo matchers and computers. Therefore, as a compromise, an execution time of around one minute was deemed to be the maximum acceptable for use in the interactive environment intended for the ASP stereo vision system.

3. *Density.* In the majority of applications, a dense set of surface measurements, and hence stereo correspondences, are required to accurately represent the surface that is to be measured. This is particularly true in many of the applications envisioned for the ASP stereo vision system where the surface to be measured (e.g. the surface of a patient's body in medical applications) is *not* a simple geometric shape that can be interpolated from a small number of measurements.

8.1.3. Shortcomings of the original MSSM stereo matcher

No stereo matcher has been reported in the computer vision literature that fulfils all three of the requirements listed in sub-section 8.1.2. The original MSSM stereo matcher that was available for use in the ASP stereo vision system was no exception. In fact, this version of the algorithm was some way from meeting any of the above requirements. However, unlike many other stereo matching algorithms, there was no inherent theoretical reason preventing these shortcomings from being overcome in the MSSM stereo matcher.

A good example of the above is that area-based stereo matchers, such as the MSSM stereo matcher, offer the potential to satisfy fully the third of the above requirements by outputting the densest possible set of stereo correspondence measurements, i.e. one

disparity measurement for every pixel in the input stereogram. (This contrasts with feature-based stereo matchers that inherently produce only a sparse set of disparity measurements). However stereo matchers such as the MSSM stereo matcher are designed to work best with stochastically based images that exhibit the optimum amount of image texture, e.g. random-dot stereograms. Unfortunately, the image statistics of most natural scenes only approximate to this optimum case. Scale-space stereo matching algorithms have therefore tended to perform less satisfactorily on non-synthetic images, particularly in regions of low image texture [Cochran, 1989; Hoff, 1989; Mohan, 1989]. Thus the effective density of usable disparity measurements produced by this class of algorithm is reduced and, if not addressed, would be unlikely to fulfil the density requirement given above.

However, vast improvements have been reported [Urquhart, 1990; Nishihara 1987] for natural scenes sparse in surface texture when they are bathed in *textured light*. It has been demonstrated that by projecting randomly textured light onto the scene under investigation, a level of accuracy approaching that achieved for random-dot stereograms is possible using the MSSM stereo matcher [Siebert, 1990]. Using this technique, it is therefore possible to exploit the full potential of area-based stereo matchers to produce dense disparity measurement even for natural scenes.

In theory, the original MSSM stereo matcher was capable of estimating disparity to approximately $^1/_{30}$ of a pixel. This theoretical precision, in combination with the original MSSM stereo matcher's maximum practical dynamic range of ±20 pixels, meant that it could provide disparity estimates over a range of approximately 1200 quantisation levels. For the 500mm depth of the viewing volume specified for the ASP stereo vision system, this would imply that an average depth measurement accuracy of roughly ±0.21mm was possible. Clearly, if this level of accuracy could be achieved in practice then the MSSM stereo matcher would certainly meet the accuracy requirement given in sub-section 8.1.2. However the actual matching accuracy achievable by the MSSM stereo matcher had never been investigated and hence it was possible that this could fall dramatically short of the theoretically predicted best performance.

However the biggest concern about the performance of the original MSSM stereo matcher was its extremely slow speed of operation. Typical execution times for the original version (using a Sun 4 workstation) were 30 minutes to match a stereo pair of 128 × 128 pixel images with ±5 pixels of disparity and 7 hours to match a stereo pair of 256 × 256 pixel images with ±12 pixels of disparity. It was a requirement of the ASP stereo vision system that full CCIR sized images (768 × 576 pixels) with ±20

pixels of disparity be matched. Clearly the original MSSM stereo matcher would have taken days to perform this and certainly not one minute as specified in the second of the requirements listed in sub-section 8.1.2.

8.1.4. Overview of MSSM development

By employing the textured illumination technique described in the previous sub-section, the MSSM stereo matcher was able to provide the required density of disparity measurements. However the shortcomings of the MSSM stereo matcher with respect to the other two requirements given in sub-section 8.1.2 had still to be addressed. The objective of the work described in this chapter was to investigate and then address the slow execution speed of the original MSSM stereo matcher, while the objective of the next chapter was to address the lack of information regarding the actual performance of the MSSM stereo matcher.

It is clear from the discussion in sub-section 8.1.3 that a speed increase of approximately 1000 times would be necessary if the MSSM stereo matcher were to meet the desired speed of execution. Although this may seem inconceivable, some early experiments performed by the author [Urquhart, 1989] had indicated that dramatic speed improvements of this order could well be possible using image pyramids.

The development of the MSSM stereo matcher took place in two stages, resulting in Versions II and III respectively. Version II of the MSSM stereo matcher was developed completely by the author and consisted of a total re-implementation of the original algorithm to use image pyramids. Version II achieved a speed increase of approximately 50 times that of the original MSSM stereo matcher. Version II was in turn completely re-implemented by Arthur van Hoff at the Turing Institute who had extensive expertise in optimising the performance of computer algorithms. Van Hoff's Version III of the MSSM stereo matcher achieved approximately a further 50 times improvement in the execution speed. All three versions are described in this chapter.

The objective of the work described in the next chapter (Chapter 9) was to address the lack of information regarding the actual performance of the MSSM stereo matcher. It was decided to delay this work until the development of the algorithm had been completed. Firstly, this would mean that the performance information obtained would be relevant to the currently operational version of the MSSM stereo matcher and not to an old redundant version. Secondly, it was realised that the original version was far too slow to allow processing of the large range of test stereo images that

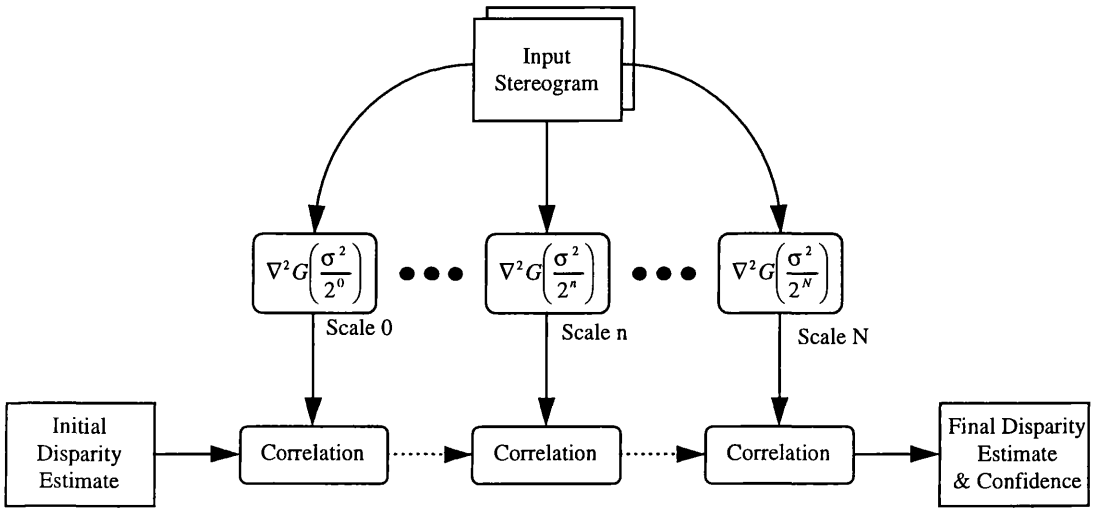


Figure 8.1. The basic structure of the MSSM stereo matcher.

would be required for a rigorous analysis of its performance. It was therefore the performance of Version III of the MSSM stereo matcher that was the subject of the analysis described in the next chapter.

8.2. Original version

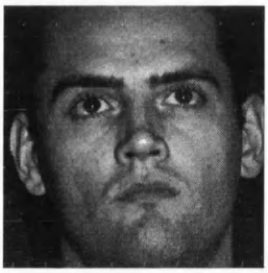
8.2.1. Basic structure

The original MSSM stereo matcher was developed by Jin Zhengping (Joseph Jin) and Peter Mowforth at the Turing Institute during the late 1980s [Zhengping, 1988; Zhengping, 1989]. The structure of the algorithm is described briefly below so that the modifications made to produce Versions II and III can be better understood. A fuller description of the original MSSM stereo matcher is given in [Urquhart, 1990].

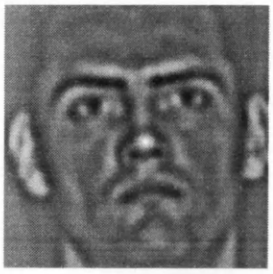
It can be seen from figure 8.1 that the basic structure of the MSSM stereo matching algorithm is two stage in nature. The first stage is known as the *blurring stage* and the second stage is known as the *matching stage*.

8.2.2. Blurring stage

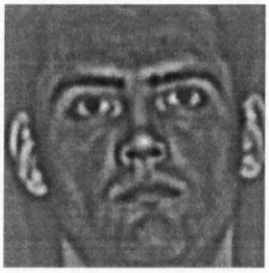
The blurring stage generates a *scale space* series of band-pass images of increasing spatial frequency by filtering the input stereo images with a series of Laplacian of Gaussian filters with reducing standard deviation. If the intensity values of the left and right input images are $I_i(x, y), i = L, R$, then the left and right band-pass images,



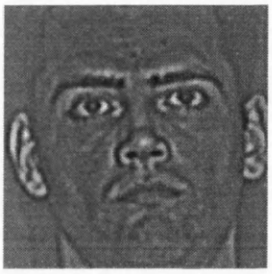
Input



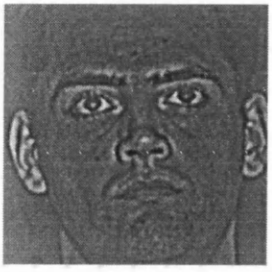
Scale 0



Scale 1



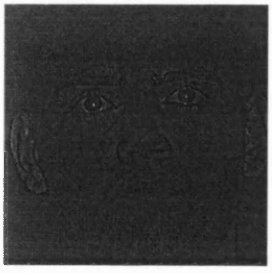
Scale 2



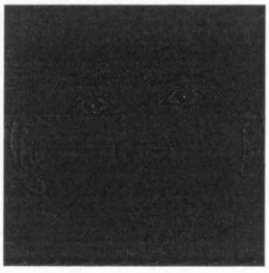
Scale 3

Scale 4

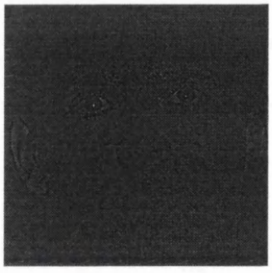
Scale 5



Scale 6



Scale 7



Scale 8

Figure 8.2. Scale space images produced by the blurring stage of the original version of the MSSM stereo matcher. (The images have been scaled for display purposes.)

$B_{i,s}(x, y)$, at each scale s are generated using the following convolution equation:

$$B_{i,s}(x, y) = \nabla^2 G_s(x, y) * I_i(x, y) = \sum_{u=-U}^{U} \sum_{v=-V}^{V} I_i(x+u, y+v) \cdot \nabla^2 G_s(u, v) \qquad (8.1)$$

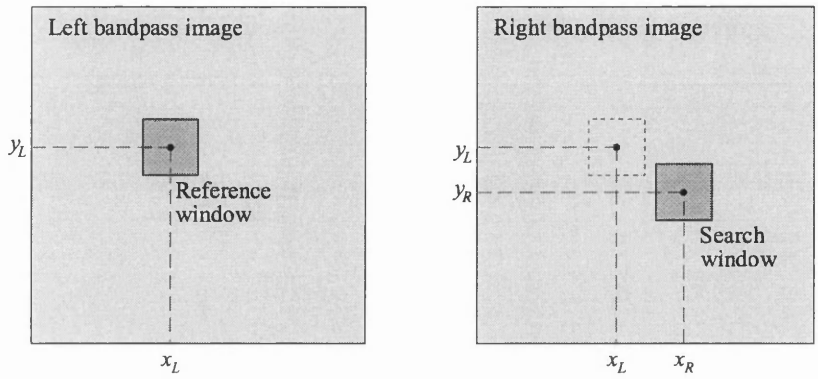


Figure 8.3. Matching is carried out between the reference window centred at (x_L, y_L) in the left band-pass image and the search window centred at (x_R, y_R) in the right band-pass image.

$\nabla^2 G_s(x, y)$, the Laplacian of Gaussian mask used at scale s, is calculated using a standard deviation, σ_s, given by:

$$\sigma_s = \sigma_0 \cdot \alpha^s \tag{8.2}$$

σ_0, the standard deviation used at scale 0, the lowest frequency band-pass, is equal to the maximum disparity expected in the input stereogram. The standard deviation used at each subsequent scale, σ_s, is then α times that of the previous scale. The theoretical limit of octave separated scales is achieved using α = 0.5, however the actual implementation used α = 0.7 to provide greater tolerance to image noise. The final, highest frequency band-pass, scale is reached when $\sigma_s \leq$. 0.34. The horizontal and vertical size, i.e. the number of columns and rows, of the Laplacian of Gaussian mask, $2U+1$ and $2V+1$ respectively in equation (8.1), are directly related to σ_s, the standard deviation at the current scale. Figure 8.2 illustrates the scale space images generated by the blurring stage of the original version of the MSSM stereo matcher.

8.2.3. Matching stage

After the blurring stage has generated the scale space series of stereo band-pass images, $B_{i,s}(x, y)$, the matching stage then matches each of them in turn, beginning at scale 0. The output of the matching process at each scale s is a pair of horizontal and vertical disparity estimates, $\Delta x_s(x_L, y_L)$ and $\Delta y_s(x_L, y_L)$ respectively, for each pixel, (x_L, y_L) in the left band-pass image for that scale. These disparity estimates are obtained by refining $\Delta x_{s-1}(x_L, y_L)$ and $\Delta y_{s-1}(x_L, y_L)$, the disparity estimated for (x_L, y_L) at the previous scale. High precision is obtained by making the disparity refinement finer at each successive scale.

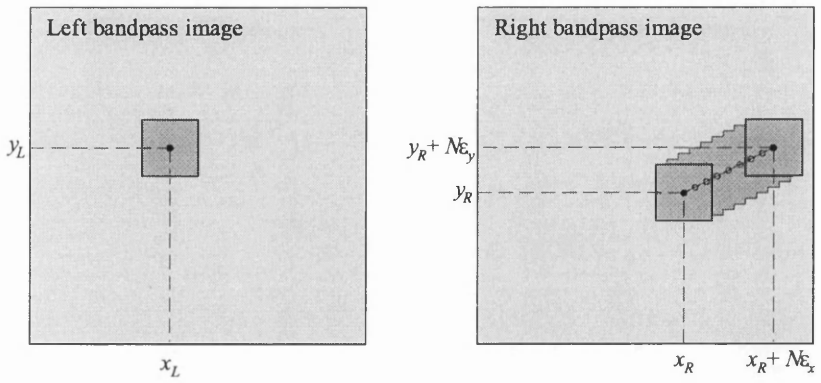


Figure 8.4. The search window is moved to a number of trial matching points in the right band-pass image.

Disparity refinement for each pixel (x_L, y_L) at each scale s, is performed by calculating the correlation between windowed regions of the left and right band-pass images as shown in figure 8.3. The *reference window* is centred at (x_L, y_L) in the left band-pass image, and the *search window* is initially centred at (x_R, y_R) in the right band-pass image, where:

$$x_R = x_L + \Delta x_{s-1}(x_L, y_L) \tag{8.3}$$

$$y_R = y_L + \Delta y_{s-1}(x_L, y_L) \tag{8.4}$$

The disparity refinement process then proceeds by moving the search window to $N+1$ *trial match points*, $(x_R + n\varepsilon_x, y_R + n\varepsilon_y), n = 0,1,\cdots,N$, in the right band-pass image as illustrated in figure 8.4, where ε_x and ε_y are the horizontal and vertical increments between the trial match points. The correlation between the reference and search windows, $\rho_{LR,s}(x_L, y_L, x_R + n\varepsilon_x, y_R + n\varepsilon_y)$, is calculated for each trial match point in turn, as described in sub-sub-section 8.2.4. This search process is terminated once a maximum has been detected in the calculated correlation values or once all $N+1$ trial match points have been visited. The disparity at the trial match point causing the maximum correlation is output as the final horizontal and vertical disparity estimates for that scale, $\Delta x_s(x_L, y_L)$ and $\Delta y_s(x_L, y_L)$ respectively. The corresponding maximum correlation value is output as the confidence value for that scale.

The horizontal and vertical search increments, ε_x and ε_y, are calculated as follows:

$$\varepsilon_x = \frac{\varepsilon \cdot E_x}{\sqrt{E_x^2 + E_y^2}} \tag{8.5}$$

$$\varepsilon_y = \frac{\varepsilon \cdot E_y}{\sqrt{E_x^2 + E_y^2}} \tag{8.6}$$

$$\varepsilon = \frac{\sigma_s}{10} \tag{8.7}$$

$$E_x = \rho_{LR,s}(x_L, y_L, x_R + \varepsilon, y_R) - \rho_{LR,s}(x_L, y_L, x_R, y_R) \tag{8.8}$$

$$E_y = \rho_{LR,s}(x_L, y_L, x_R, y_R + \varepsilon) - \rho_{LR,s}(x_L, y_L, x_R, y_R) \tag{8.9}$$

8.2.4. Windowed correlation calculation

Correlation between the reference window situated at (x_L, y_L) in the left band-pass image and the search window situated at a trial match point (x_R, y_R) in the right band-pass images is calculated as follows:

$$\rho_{LR,s}(x_L, y_L, x_R, y_R) = \frac{C_{LR,s}(x_L, y_L, x_R, y_R)}{\sqrt{\sigma_{L,s}^2(x_L, y_L)\sigma_{R,s}^2(x_R, y_R)}} \tag{8.10}$$

$$\sigma_{L,s}^2(x_L, y_L) = \sum_{u=-U}^{U}\sum_{v=-V}^{V} G_s(u,v) \cdot B_{L,s}(x_L + u, y_L + u) \cdot B_{L,s}(x_L + u, y_L + u) \tag{8.11}$$

$$\sigma_{R,s}^2(x_R, y_R) = \sum_{u=-U}^{U}\sum_{v=-V}^{V} G_s(u,v) \cdot B_{R,s}(x_R + u, y_R + u) \cdot B_{R,s}(x_R + u, y_R + u) \tag{8.12}$$

$$C_{LR,s}(x_L, y_L, x_R, y_R) = \sum_{u=-U}^{U}\sum_{v=-V}^{V} G_s(u,v) \cdot B_{L,s}(x_L + u, y_L + u) \cdot B_{R,s}(x_R + u, y_R + u) \tag{8.13}$$

Where $\sigma_{L,s}^2(x_L, y_L)$ is the variance of the region of the left band-pass image within the reference window, $\sigma_{R,s}^2(x_R, y_R)$ is the variance of the right band-pass image within the search window and $C_{LR,s}(x_L, y_L, x_R, y_R)$ is the covariance of the reference and search windows (a mean of zero is assumed). $G_s(u,v)$ is a two dimensional Gaussian function that is applied to the reference and search windows to weight the statistics calculations to the centre of the windows. The standard deviation of the Gaussian weighting function used at each scale is 1.3 times σ_s, the standard deviation used in the blurring stage for that scale.

It should be noted that the centre of the search window in the right band-pass image, (x_R, y_R), will in general be at a sub-pixel location. Interpolation is therefore required to derive the values of $B_{R,s}(x_R + u, y_R + u), u = -U, \cdots, U, v = -V, \cdots, V$ contained

within the search window during the calculation of $\sigma^2_{R,s}(x_R, y_R)$ and $R_{LR,s}(x_L, y_L, x_R, y_R)$.

8.2.5. Two dimensional sub-pixel interpolation

The original MSSM stereo matcher calculates the intensity value of an image at a sub-pixel location (x, y) by interpolating the intensity values at the four nearest integer pixel locations $F(X,Y)$, $F(X+1,Y)$, $F(X,Y+1)$ and $F(X+1,Y+1)$ as follows:

$$\begin{aligned} F(x,y) = &(1-\lambda_X)(1-\lambda_Y)F(X,Y) + \lambda_X(1-\lambda_Y)F(X+1,Y) \\ &+(1-\lambda_X)\lambda_Y F(X,Y+1) + \lambda_X\lambda_Y F(X+1,Y+1) \end{aligned} \tag{8.14}$$

$$X = \lfloor x \rfloor, \quad Y = \lfloor y \rfloor \tag{8.15}$$

$$\lambda_X = x - X, \quad \lambda_Y = y - Y \tag{8.16}$$

8.2.6. Implementation

The original MSSM stereo matcher was written in the C programming language using the HIPS image processing library [Landy, 1984]. HIPS provides both an extensive library of common image processing tools and a ready framework for the development of new image processing modules, such as MSSM.

8.3. Version II

8.3.1. Irregular pyramid implementation

Version II of the MSSM stereo matcher was developed by the author at the Turing Institute in 1992 and was based upon the results of an earlier investigation that is reported in [Urquhart, 1989]. The motivation for this re-implementation of the algorithm was that approximately 99.8% of the execution time for the original version of the MSSM stereo matcher was spent in the matching stage. The majority of this time was spent matching the lowest frequency band-pass images due to the extremely large size of the reference and search windows used at the lowest scales. This is illustrated in table 8.1 which shows the size of the reference and search windows at each scale for an initial blurring sigma of 5.000.

In [Burt, 1983; Burt, 1988] Burt defines a Gaussian pyramid in which an input image is convolved with a series of Gaussian filters whose size, i.e. sigma, successively doubles. This has the effect of low-pass filtering the images with a band limit that is

successively reduced in octave steps. As the band limit reduces in octave steps, so the sample density of the image may also be decreased by a factor of two in each dimension with virtually no loss of information. This results in a "pyramid" of images in which each image has half the number of rows and columns of the one below it.

Scale	0	1	2	3	4	5	6	7	8
Sigma	5.000	3.500	2.450	1.715	1.200	0.840	0.558	0.411	0.340
Columns	39	27	19	13	9	7	5	3	3
Rows	19	13	9	7	5	3	3	1	1

Table 8.1. Blurring sigma and matching window rows and columns for each scale in the original MSSM stereo matcher.

There is a clear similarity between the series of Gaussian filters used to generate Burt's Gaussian pyramid and the series of Laplacian of Gaussian filters used in the blurring stage of the original MSSM stereo matcher. The aim was therefore to apply Burt's idea of successive image re-sampling in Version II of the MSSM stereo matcher so that the lowest scale (i.e. lowest frequency) band-pass images would be the most reduced in size. This would result in a dramatic speed up not only because the reduced number of pixels in the re-sampled images require correspondingly fewer searches, but also because the windows used during matching can be reduced in size by the same amount as the images themselves.

However, two significant differences were evident between the generation of Burt's Gaussian pyramid and the blurring stage of the MSSM stereo matcher. Firstly, Laplacian of Gaussian filters are used as opposed to Gaussian filters, and secondly, the sigma values of successive filters are reduced by a factor of 0.7 and not 0.5 as required for octave separation. In order to resolve these differences it was necessary to refer to some basic sampling theory. This tells us that the maximum spatial frequency that can be represented in an image is 0.5 cycles/pixel (π rad/pixel) and that the image size can be re-sampled by a factor of two without aliasing if it is first low-pass filtered with a cut-off of 0.25 cycles/pixel ($\pi/2$ rad/pixel), i.e. half the original maximum spatial frequency. The allowable sub-sampling factor, S, can therefore be expressed as the following function of the cut-off frequency of the low-pass filter, $\omega_{cut\text{-}off}$ (rad/pixel):

$$S = \pi/\omega_{cut-off} \tag{8.17}$$

Laplacian of Gaussian filters, in common with all practical filters, do not have a "brick wall" cut-off and an aliasing component will remain whatever cut-off frequency is chosen. A compromise solution was therefore required that would ensure an acceptably small amount of aliasing while still allowing a reasonable amount of sub-sampling. It was decided that an acceptable compromise could only be determined by experimentation. The upper amplitude half sensitivity point, ω_{upper}, of the Laplacian of Gaussian filter was chosen as the initial cut-off frequency to determine how much re-sampling could be performed at each scale in the MSSM stereo matcher. [Zhengping, 1988] shows that ω_{upper} can be calculated from the sigma for that scale, σ_s, using the following equation:

$$\omega_{upper} = 2.314452973/\sigma_s \text{ (rad/pixel)} \tag{8.18}$$

Therefore, the sub-sampling factor, S_s, chosen for use at each scale was:

$$S_s = \frac{\pi\sigma_s}{2.314452973} \tag{8.19}$$

It was decided to implement the image pyramid using only image reductions in powers of two. Although this was a rather conservative approach, it was chosen for the simplicity with which it could be implemented, analysed and debugged. Consequently, the sub-sampling factor used at each level in the pyramid was rounded down to the nearest power of two as indicated in table 8.2 below. Figure 8.5 shows the irregular pyramid structure of the images that results and figure 8.6 shows an example of such an irregular pyramid.

Scale	0	1	2	3	4	5	6	7	8
σ_s	5.000	3.500	2.450	1.715	1.200	0.840	0.558	0.411	0.340
S_s	6.786	4.751	3.326	2.328	1.629	1.140	0.757	0.558	0.462
Reduction	4	4	2	2	1	1	1	1	1

Table 8.2. Sub-sampling factor and actual image reduction factor applied for a scale space sequence with an initial sigma of 5.0.

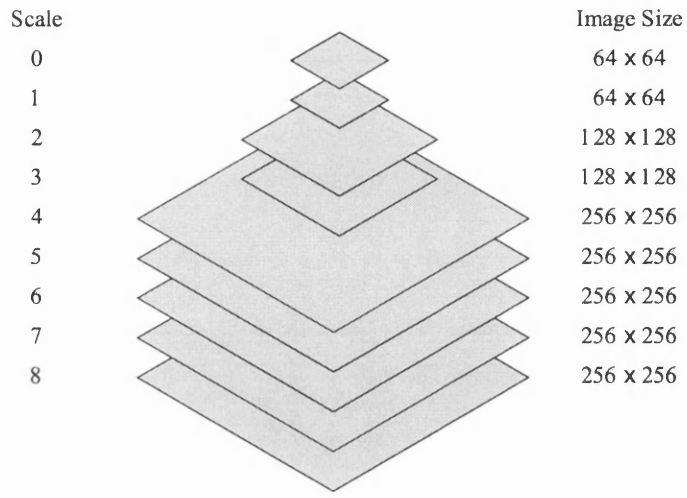


Figure 8.5. The irregular pyramid structure used in Version II of the MSSM stereo matcher.

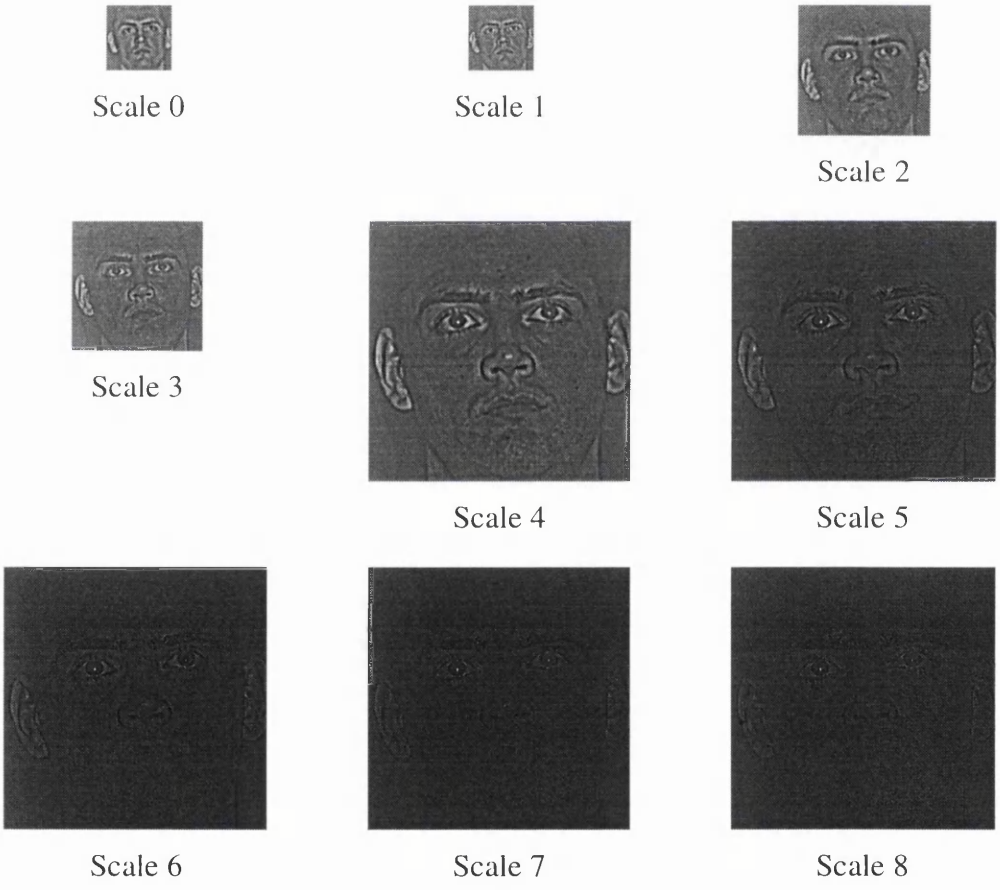


Figure 8.6. Irregular pyramid of images produced by the blurring stage of the Version II of the MSSM stereo matcher from the same input image as shown in figure 8.2. (The images have been scaled for display purposes.)

Whenever a disparity map was to be passed between scales with different reduction factors, it was first necessary to expand it up by a factor of two to the size of the images for the later scale. This was done by pixel replication using bilinear interpolation. It was also necessary to multiply the actual disparity values by two since a disparity shift of one pixel in the smaller image would represent a shift of two pixels in the larger image. Limiting the reduction factors to powers of two ensured that any expansion required between scales would always be an integer.

8.3.2. Complete re-coding

The constant revisions and modifications that had been made to the source code during the research and development of the original version of the MSSM stereo matcher had meant that it had become quite untidy and very difficult to follow. Extensive additional re-structuring would have been required in order to incorporate the irregular image pyramid structure that was described in the previous sub-sub-section. Furthermore, extensive modification to the source code was required to make MSSM compatible with the second version of HIPS [Cohen 1991, Landy 1991], which had just become available and which was substantially different from the original. Therefore, it was decided to completely re-write the source code for Version II of the MSSM stereo matcher to commercial coding standards. In the course of doing this, further smaller speed improvements were achieved by improving the efficiency of parts of the original code.

8.3.3. Results

Examples of the speed improvement obtained with Version II of the MSSM stereo matcher are given in table 8.3. No detailed comparison of the matching performance of the two versions of the algorithm was deemed to be necessary since it was observed that the disparity maps output by the two versions were virtually identical for a range of input stereograms.

Image size	Disparity range	Original time	Version II time
128×128	± 5 pixels	2055s	119s
256×256	± 12 pixels	25086s	500s

Table 8.3. Comparison of execution times (on a Sun 4 workstation) for the original and Version II MSSM stereo matchers.

It is notable that the speed improvement obtained with Version II of the MSSM stereo matcher is greatest for the earliest scales in the scale space series since these result in the largest image reductions. In fact, the time required to match the earliest scales was reduced by so much using the irregular pyramid structure that it was almost insignificant in comparison with that required for the later full resolution scales. An important implication of this is that the large processing overhead associated with matching large disparities using the original MSSM stereo matcher no longer applies in Version II.

8.4. Version III

8.4.1. Overview

The successful development of Version II of the MSSM stereo matcher prompted an almost immediate attempt to achieve even greater performance improvements. The improved quality and documentation of the source code meant that the process of understanding exactly how the algorithm worked was much more straightforward. This allowed the development of Version III of the MSSM stereo matcher to be passed over to Arthur van Hoff at the Turing Institute with very little difficulty and allowed the author to concentrate on other aspects of the ASP project.

8.4.2. Implementation

The following major modifications were made to Version II of the MSSM stereo matcher (a more detailed description of Version III of the MSSM stereo matcher is given in [Van Hoff, 1992a; Van Hoff, 1992b]):

- The source code was once again completely rewritten in order to accommodate more easily the major changes that were made to the algorithm.
- Single precision floating point numbers were used instead of double precision numbers, resulting in a 50% reduction in the memory requirements and a speed increase of 10% without any noticeable loss in accuracy.
- Encouraged by the performance improvement obtained by the use of the irregular pyramid structure, a much less conservative regular pyramid structure, more similar to Burt's original "Laplacian pyramid" [Burt, 1983], was developed.

- A much more efficient search strategy was adopted, that required the image correlation to be calculated at approximately only a quarter of the number of locations required by Version II.

- Careful selection of the trial match points ensured that a more efficient method for sub-pixel interpolation could also be developed.

8.4.3. Results

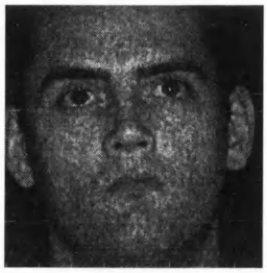
These major improvements to the MSSM algorithm, together with more minor improvements made during the implementation of Version III of the MSSM stereo matcher resulted in a further dramatic improvement in the performance of the algorithm. Although no detailed comparison of the accuracy of Version III and Version II was carried out, the output disparity maps were observed to be qualitatively very similar for a range of input stereograms. Examples of the output from Version III of the MSSM stereo matcher for a texture projected and naturally illuminated stereogram are shown in figure 8.7 and figure 8.8 respectively.

Example execution times for Version III of the MSSM stereo matcher on various computers and for a range of image sizes are shown in table 8.4. It is clear from table 8.4 that the desired performance of one minute to process a full CCIR 576 × 768 pixel image was now achievable using even a mid-range workstation such as a single processor Sun SPARCstation 10 model 41.

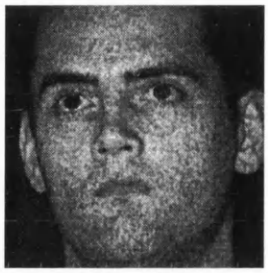
Image size (pixels)	128 × 128	256 × 256	512 × 512	576 × 768
Sun SPARCstation LX	8.4s	32.7s	130.9s	215.0s
IBM RS6000 320H	3.8s	15.5s	64.5s	104.6s
Sun SPARCstation 10	2.4s	10.7s	50.5s	71.2s

Table 8.4. Example execution times for Version III of the MSSM stereo matcher on various computers and for a range of image sizes.

It should also be noted that Version III of the MSSM stereo matcher could also operate in a one-dimensional matching mode in which only the horizontal component of disparity is estimated. This was approximately 40% quicker than the two-dimensional version shown in table 8.4.

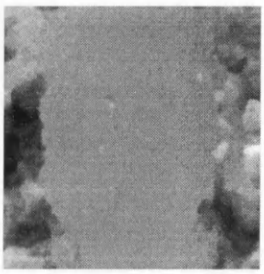


Left input



Right input

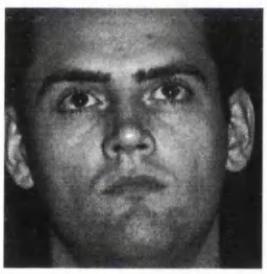
Horizontal disparity



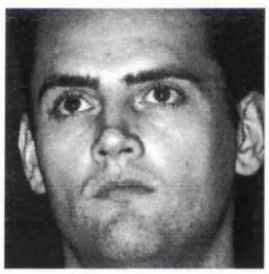
Vertical disparity

Confidence

Figure 8.7. Output from Version III of the MSSM stereo matcher for a texture projected input stereogram. (The disparity and confidence images have been scaled for display purposes.)

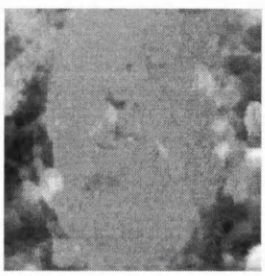


Left input



Right input

Horizontal disparity



Vertical disparity

Confidence

Figure 8.8. Output from Version III of the MSSM stereo matcher for a naturally illuminated input stereogram. (The disparity and confidence images have been scaled for display purposes.)

8.5. Conclusions

The objective of the work described in this chapter was to identify and address the deficiencies of the MSSM stereo matcher with regard to the requirements of the ASP stereo vision system. The primary deficiency was the extremely slow speed of execution of the original MSSM stereo matcher. The development programme for the MSSM stereo matcher that was described in this chapter has clearly proven to be extremely effective: resulting in an increase in the intrinsic speed of the algorithm's execution by more than 1000 times.

The contributions made at each stage of the development of the MSSM stereo matcher were significant to the overall success. The original version, developed by Jin Zhengping (Joseph Jin), set the overall structure and developed the basic theory of the algorithm. A dramatic improvement in the execution speed was achieved in Version II of the algorithm, that was developed by the author. This established a framework for highly efficient scale space pyramid structures and demonstrated their feasibility by implementing an irregular image pyramid structure. A further dramatic improvement resulted from the development of Version III of the algorithm by Arthur van Hoff. This was achieved by developing a much less conservative regular pyramid image structure and completely optimised the basic performance of the MSSM algorithm.

The development of Version III meant that the MSSM stereo matcher could now fulfil both the density and speed requirements that were given in sub-section 8.1.2. Not only could it still produce the required dense set of stereo correspondence measurements, but it could also now process a full resolution CCIR image in the required time of approximately one minute. This would previously have required days of processing with the original MSSM stereo matcher. Additionally, the highly optimised 2D interpolation scheme used in Version III of the stereo matcher resulted in a substantial reduction in the difference in time required for one and two dimensional matching. The time required for two dimensional matching was now no more than twice that required for one dimensional matching. Furthermore, the pyramid image structure that was adopted in Versions II and III of the MSSM stereo matcher meant that there was no longer a considerable processing overhead associated with matching stereograms containing large amounts of disparity. The dynamic range of matching was therefore increased to approximately ±50 pixels for CCIR resolution images.

The ASP sensor head configuration derived in chapter 6 had been designed specifically to minimise the amount of vertical disparity and hence facilitate matching in only one dimension. It had also been designed to restrict horizontal disparity to the range ±20 pixels, in accordance with the dynamic range of the original version of the

MSSM stereo matcher. As a consequence of the improvements made to the MSSM stereo matcher, both in its dynamic range and in the relative speed of two dimensional matching, the originally specified requirements of the ASP sensor system could be relaxed. It was therefore feasible to extend the operation of the ASP sensor system, beyond the original specification, to fully utilise the improvements made to the MSSM stereo matcher. For example, a viewing volume much closer than the originally specified mean object distance of 5.0m could now be successfully processed.

Finally, Version III of the MSSM stereo matcher retained the basic capability of the algorithm to generate disparity estimates with sub-pixel *resolution*, by means of various stages of sub-pixel image interpolation. However the *accuracy* of the MSSM stereo matcher, and in particular its ability to meet the remaining accuracy requirement of the ASP stereo vision, was still unknown. Indeed, its accuracy became even more uncertain with the development of Version III of the algorithm. The final disparity step size of $^1/_{30}$ of a pixel set a theoretical maximum accuracy for the original version of the MSSM stereo matcher that was retained in Version II. However this was not retained in Version III and consequently there was no longer a simple basis for estimating its actual matching accuracy. Consequently a series of experiments were designed with the intention of addressing this remaining shortcoming of the MSSM stereo matcher. These experiments are described in chapter 9.

Chapter 9

Stereo matcher accuracy analysis

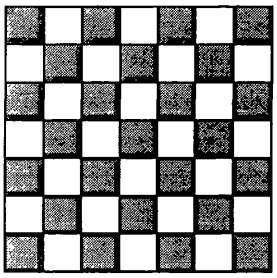
9.1. Introduction

9.1.1. Objective

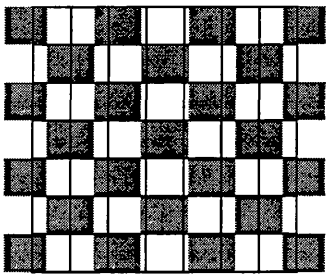
The development of Version III of the MSSM stereo matcher that was described in the previous chapter addressed its slow speed of execution. The final remaining shortcoming of the algorithm was therefore the lack of information regarding its performance in general and its matching accuracy in particular. The objective of the experiments described in this chapter was therefore to address this shortcoming by analysing the performance of Version III of the MSSM stereo matcher.

Although the MSSM stereo matcher has been designed specifically so that disparity can be measured with sub-pixel precision, the actual accuracy of the disparity measurements it generates had never been tested. This situation was exacerbated with the development of Version III of the MSSM stereo matcher since not even its theoretical maximum accuracy was known. It was therefore uncertain whether the MSSM stereo matcher would fulfil the accuracy requirements of the ASP stereo vision system.

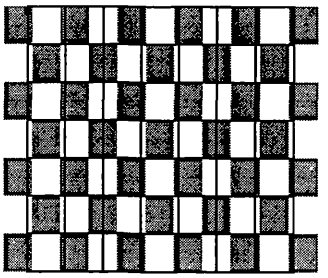
Consequently it was a pressing requirement that the accuracy of Version III of the MSSM stereo matcher be measured so that its suitability for use in the ASP stereo vision system, or in any other application for that matter, could be gauged. The objective of the experiments described in this chapter was therefore to characterise the matching accuracy of Version III of the MSSM stereo matcher and hence to verify its suitability for use in the ASP stereo vision system.



Original checker-board image.



Disparity gradient of +0.25.



Disparity gradient of −0.25.

Figure 9.1. Illustration of the stretching/shrinking effect of disparity gradients of +0.25 and −0.25 on a 7 × 7 pixel checkerboard image.

9.1.2. Factors affecting stereo matcher accuracy

Experience acquired using all versions of the MSSM stereo matcher has shown that the primary factor affecting its accuracy is insufficient surface texture in the input imagery. However, in the case of the ASP stereo vision system, textured scene illumination has been successfully developed to address this problem [Urquhart, 1990]. Therefore, the relationship between surface texture and stereo matcher accuracy was not under investigation in this chapter.

For a given level of surface texture, and ignoring all other factors, the MSSM stereo matcher will exhibit a base level of matching accuracy. However, the accuracy will fall below this base figure whenever differences occur between the images in a stereo-pair. Therefore, since images captured from different viewpoints will always contain differences (such as those induced by the very disparity that the matcher is trying to measure!), the accuracy achieved by the MSSM stereo matcher will, in practice, always be worse than the base level. The objective of this chapter was to characterise the effect on match accuracy of the following sources of differences between the images in a stereo pair:

1. A uniform disparity offset between the images. This will result in a translation between the left and right images. (Although this should not effect the accuracy of the MSSM stereo matcher, it had not previously been tested with any rigour).

2. A changing disparity, or a *disparity gradient*, which will result in one image being either stretched or shrunk with respect to the other. This stretching/shrinking effect of a uniform disparity gradient is illustrated in figure 9.1.

3. Random noise injected by the imaging sensor which will be different in either image.

Two other mechanisms exist by which differences can occur between the images in a stereo pair: occlusion and specular reflection. An occlusion is a region of one image that is completely unseen in the other. There is no correct match in this situation, and the result is therefore complete match failure in the region of the occlusion (as opposed to an area of reduced match accuracy). Similarly, complete match failure also occurs in the vicinity of a specular reflection if it appears in only one of a stereo pair of images.

A quite different effect is often observed when a specular reflection appears in both images of a stereo pair. In this case, the specular reflection generally presents a very strong local match feature. However, because the image location of the specular reflection is dependent on the relative positions of both the light source and the camera, the perceived location of the specular reflection is not actually on the physical surface that is being imaged. Therefore, the result is a actually a strong "false" match. Although the effects of both occlusion and specular reflection were deemed to be outside the scope of the work described in this chapter, it should be noted that the author has extensively investigated the process by which occlusions occur and possible methods for detecting them [Urquhart, 1990].

9.1.3. Approach

The objective of this chapter was to characterise the performance envelope of the MSSM stereo matcher. The adopted approach was to investigate how the factors described in sub-section 9.1.2 individually affect the accuracy of the MSSM stereo matcher. These factors were: the amount of disparity that is implicit in the stereograms, the rate of change (or gradient) of the disparity and the signal-to-noise ratio of the images. In addition, it was postulated that the behaviour of the MSSM stereo matcher was different when presented with exact integer disparities than when presented with non-integer sub-pixel disparities. It was therefore decided to investigate these two cases separately.

It was intended that this investigation would result in a basic characterisation of the accuracy of the stereo matcher, albeit "under laboratory conditions". The vast differences in the image characteristics of stereograms of real world scenes meant that it was unlikely that any absolute figure for match accuracy derived from these experiments would be directly applicable to all individual cases "outside the laboratory". However, it was intended to obtain sufficient characterisation information

to determine the suitability of the MSSM stereo matcher for use in the controlled context of the ASP stereo vision system. It also was anticipated that the experiments would establish a basic performance envelope for the MSSM stereo matcher and determine the relative effect of each of the basic factors listed in the previous section.

9.2. Generation of test stereograms

In order to perform the planned analysis of the MSSM stereo matching algorithm, it was first necessary to provide stereograms with easily controllable characteristics. In order to determine the accuracy with which a given stereogram had been matched, it was also necessary to know accurately the underlying disparity in the stereogram. The error in disparity maps output from MSSM could then be determined by subtracting them from the known disparity. It was decided that the only practical method of providing the large number of such stereo images that would be required for the analysis was to synthetically generate stereograms similar to that shown in figure 9.2.

9.2.1. Disparity map synthesis

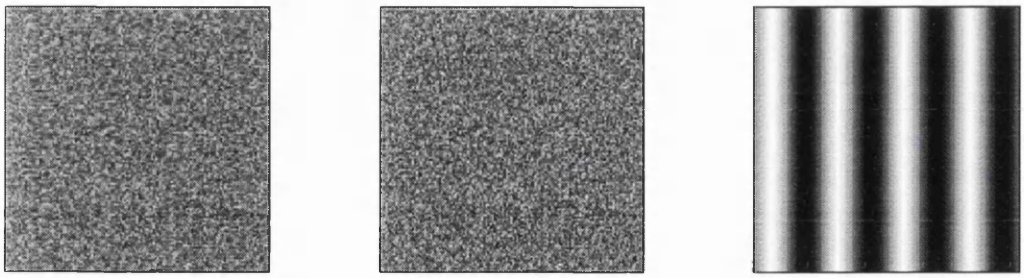
The random noise stereograms used as the input to each experiment were constructed by first synthesising the required (one-dimensional) disparity map. This was done using the standard HIPS-2 filter `fcalcpix` which allows a floating point format image to be generated with pixel values assigned using C code statements specified on the command line. For example, the disparity map shown in figure 9.2 was synthesised using the following command:

```
fcalcpix -c 128 128 -s "opix = sin ((double) c * H_PI / 16)" \
      > disp.hips
```

Where `opix` is the value of the output pixel, `c` is the pixel's column position and `H_PI` is π. Therefore the resulting HIPS image, `disp.hips`, is a 128×128 pixel horizontal sinusoid of amplitude 1.0 pixels and period 32 pixels.

9.2.2. Generation of right image

The next stage in the construction of the test stereograms was to synthesise one of the stereo pair of images. The other image would then be generated by image warping the first via the disparity map. It will become clear later that it was simpler to synthesise the right image first and then to generate the left from it by image warping.



Left image. *Right image.* *Disparity map.*

Figure 9.2. Example of a random noise stereogram and its disparity map.

A random noise image was chosen for use as the right image because this would both ensure sufficient texture for matching and allow many different uncorrelated examples to be generated easily. Ideally, white noise would provide the best possible input to the MSSM stereo matcher because it has the optimum auto-correlation function (1.0 for no offset and zero elsewhere). However it was not possible to generate easily white noise using the standard library of HIPS filters. The closest alternative was to use the HIPS `gnoise` filter to generate Gaussian noise images of high standard deviation. (As the standard deviation of a Gaussian noise signal tends to infinity, so the signal tends to white noise).

An additional factor favouring the use of Gaussian noise images was that the method of texture projection to be employed in the ASP system was designed to produce images that resembled, as closely as possible, Gaussian noise images. It was therefore expected that an approximate indication of the expected performance of the MSSM stereo matcher in the context of the ASP system might be drawn from the results obtained using synthesised Gaussian noise images.

Therefore, the next stage in the generation of the test stereograms was to synthesise a Gaussian random noise image to be used as the right image in the stereo pair. This was done using a command line such as:

```
genframe -s 128 128 -g 128 | gnoise -p 32 0 > right.hips
```

Where the HIPS-2 command `genframe -s 128 128 -g 128` generates a 128 × 128 byte format image with a uniform grey-level of 128. `gnoise -p 32 0` then adds Gaussian noise of standard deviation 32 grey-levels and 0 as the seed for the random number generator to the image. A standard deviation of 32 was used because unacceptable levels of grey-level clipping were found to occur for higher values. (Pixels with a grey-level of less than 0 or greater than 255 will be clipped. For a mean grey-value of 128 and a standard deviation of 32, this corresponds to approximately

$\pm 4\sigma$, or 0.006% of pixels.) An example of the resulting random noise image, `right.hips`, is illustrated in figure 9.2.

9.2.3. Generation of left image

The left image of the stereo pair, `left.hips`, must now be generated by image warping the right image, `right.hips`, by the disparity map, `disp.hips`. The disparity map represents the *forward mapping* from the left image to the right, so that for every pixel location (x_L, y_L) in the left image, the corresponding point in the right image (x_R, y_R) is given by:

$$x_R = x_L + \Delta x(x_L, y_L) \tag{9.1}$$

$$y_R = y_L + \Delta y(x_L, y_L) \tag{9.2}$$

Where $\Delta x(x_L, y_L)$ and $\Delta y(x_L, y_L)$ are the horizontal and vertical disparity values for (x_L, y_L). Although the left image position (x_L, y_L) is an exact integer pixel location, the right image position, (x_R, y_R), will not be because $\Delta x(x_L, y_L)$ and $\Delta y(x_L, y_L)$ are floating point values. Interpolation is therefore required to synthesise the right image intensity values at integer pixel locations, $I_R(x, y)$, from the left image, $I_L(x, y)$, and the disparity map, $\Delta x(x, y)$ and $\Delta y(x, y)$. However this forward mapping from the left image to right is fairly complex to calculate. It is in fact much simpler to perform the inverse *backward mapping*, $W_{L \leftarrow R}(\bullet)$, from the right image and the disparity map to the left image as follows:

$$\begin{aligned} I_L(x, y) &= W_{L \leftarrow R}\big(I_R(x, y), \Delta x(x, y), \Delta y(x, y)\big) \\ &= I_R\big(x + \Delta x(x, y), y + \Delta y(x, y)\big) \end{aligned} \tag{9.3}$$

The intensity value at the sub-pixel location $\big(x + \Delta x(x, y), y + \Delta y(x, y)\big)$ in the right image can be calculated quite easily by interpolating the intensity values of the nearest four pixels in the right image array using a method such as the one described in sub-section 8.2.5. Exactly this method is used in the custom HIPS-2 filter `redistort`, which was developed by Joseph Jin at the Turing Institute to perform backward image warping. The following command line was therefore used to generate the left image in the random noise stereograms from the right image and the disparity map:

```
redistort disp.hips right.hips > left.hips
```

The final step in the generation of the random noise stereograms was to combine the left and right images into a single HIPS sequence using the standard HIPS-2 `catframes` command:

```
catframes left.hips right.hips > stereo.hips
```

9.3. Basic performance tests

9.3.1. Objective

The objective of this experiment was to verify the basic performance of the MSSM stereo matcher. This was to be done by comparing the output of the MSSM stereo matcher with the ideal expected output for some standard examples of input stereograms. It was intended that this experiment would also provide some benchmark figures against which the accuracy obtained by the MSSM stereo matcher in the subsequent experiments could be judged.

9.3.2. Input stereograms

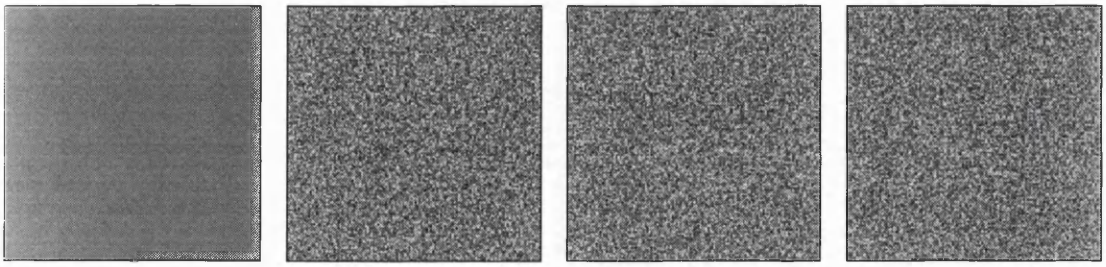
The following stereograms were chosen for analysis in this experiment because the ideal output expected from the stereo matcher was easily specified for each:

Stereogram 1: two identical images of a single grey-level. This stereogram cannot be matched at all since there is no image texture in the images. Ideally, a stereo matcher should reflect this by outputting zero disparity and minimum confidence for all pixels.

Stereogram 2: two identical random noise images. As described in sub-section 9.2.2, a stereogram consisting of two identical white noise images would have provided the ideal input to this experiment. The ideal output expected in this case would be zero disparity and maximum confidence for all pixels. However, for the reasons indicated in sub-section 9.2.2, Gaussian noise images with a high standard deviation were substituted for ideal white noise images.

Stereogram 3: a random noise image and its inverse image. The cross-correlation function for an ideal white noise signal of infinite extent and its inverse is -1.0 for no offset and zero otherwise. The ideal output from a stereo matcher for this input would therefore be zero disparity and minimum confidence for all pixels. However, as for Stereogram 2, Gaussian noise images were used instead of the ideal white noise images.

Stereogram 4: two different random noise images. The cross-correlation function between two different ideal white noise signals of infinite extent is zero for all offsets. The ideal output from a stereo matcher for this input would therefore be zero disparity



`plane.hips` `noise1.hips` `inv_noise1.hips` `noise2.hips`

Figure 9.3. The input images for the basic performance tests.

and minimum confidence for all pixels. Again, Gaussian noise images were used instead of the ideal white noise images.

9.3.3. Method

Four images were generated to form the input stereograms for the experiment: `plane.hips`, an image with the intensity value of every pixel set to 128; `noise1.hips`, a Gaussian random noise image; `inv_noise1.hips`, the inverse image of `noise1.hips`; and `noise2.hips` a Gaussian random noise image generated with a different random number seed from `noise1.hips`. A standard deviation of ±32 grey-levels was used to generate the Gaussian noise images since an unacceptable amount of grey-level clipping was found to occur with higher standard deviations. The four input images are illustrated in figure 9.3.

Stereogram	Left image	Right image
1	`plane.hips`	`plane.hips`
2	`noise1.hips`	`noise1.hips`
3	`noise1.hips`	`inv_noise1.hips`
4	`noise1.hips`	`noise2.hips`

Table 9.1. Combinations of left and right images used in the basic performance test.

The required stereograms were constructed by combining two of the above images: one to form the left image and one to form the right as indicated in table 9.1. These

stereograms were then input to the MSSM stereo matcher and the resulting disparity and confidence maps were recorded.

9.3.4. Results

The disparity and confidence maps generated by the MSSM stereo matcher from each of the four image combinations indicated in table 9.1 are shown in figure 9.4. The mean and standard deviations of the resultant disparity and confidence maps are given in table 9.2.

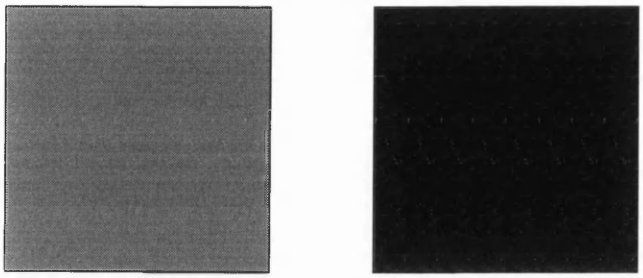
	Disparity (pixels)		Confidence	
Stereogram	Mean	Std. dev.	Mean	Std. dev.
1	0.000	±0.000	0.000	±0.000
2	0.000	±0.069	0.762	±0.006
3	0.000	±0.000	0.000	±0.000
4	-0.997	±2.244	0.046	±0.096

Table 9.2. Mean and standard deviations of the resultant disparity and confidence maps from the basic performance tests.

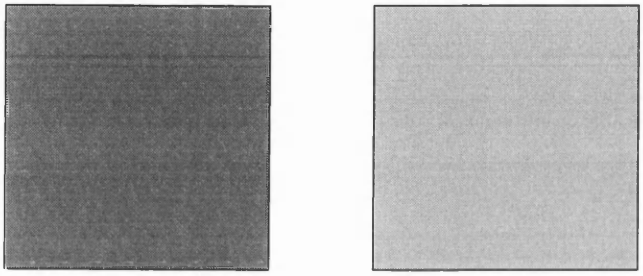
9.3.5. Interpretation of results

The output of the MSSM stereo matcher for Stereogram 1 (`plane.hips` matched against itself) conforms exactly to the ideal of zero disparity and confidence for all pixels. This occurs because the input signals are uniformly zero and consequently the correlation coefficient computed between any two windows within the left and right images will also be zero at every scale of matching.

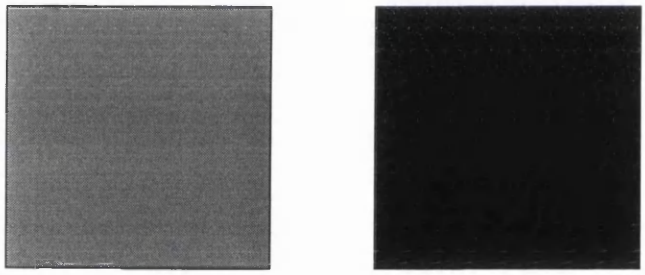
The disparity map output by the MSSM stereo matcher for Stereogram 2 (`noise1.hips` matched against itself) was found to have an RMSE of ±0.069 pixels. The slight inaccuracy observed in the output disparity map is probably due mainly to the fact that the input consisted of Gaussian noise images rather than the ideal case of white noise. The finite size of the reference and search windows used in the MSSM stereo matcher may also contribute to the observed inaccuracy, since even a



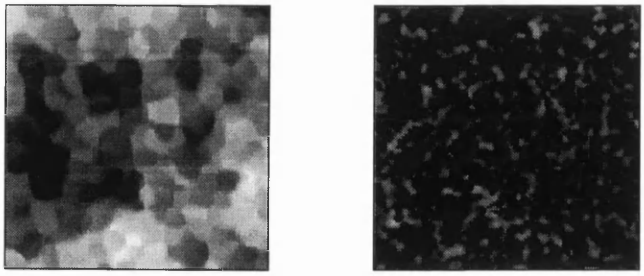
Stereogram 1.



Stereogram 2.



Stereogram 3.



Stereogram 4.

Figure 9.4. The output disparity maps (left) and confidence maps (right) from the basic performance tests. Note that, for display purposes, the disparity maps have been scaled from -5.0 (black) to +5.0 (white) and the confidence maps have been scaled from 0.0 (black) to 1.0 (white).

windowed region of a white noise signal will exhibit a sub-optimal correlation function. The RMS disparity error figure of ±0.069 pixels for this close-to-optimal case is a useful benchmark against which the algorithm's performance can be measured, e.g. in the other experiments described in this chapter.

The output confidence for Stereogram 2 was found to be consistently high with a computed mean of 0.762 and a standard deviation of ±0.006. However this is clearly significantly less than a figure of around 1.0 which might have been expected for the close-to-ideal nature of Stereogram 2. In Version III of the MSSM stereo matcher, the final confidence value output for a particular pixel is not a single correlation value. Instead, the output is calculated by summing the weighted maximum correlation for the pixel at each scale of matching. After further investigation and a careful study of the source code for Version III of the MSSM stereo matcher, a bug was found in the correlation weighting scheme that accounted for the reduced confidence values. The mean of the "raw" correlation values for Stereogram 2 were found to be 1.00, 0.99, 0.99, 0.98 and 0.98 for successive scales of matching.

The output confidence and disparity for Stereogram 3 (`noise1.hips` matched against its inverse image `inv_noise1.hips`) are exactly zero for all pixels. This indicates that the MSSM stereo matcher has, quite correctly, determined that no plausible matches exist between the left and right images. This is perhaps somewhat surprising given the sub-optimal nature of the input imagery and the scale-space matching strategy employed. However, the MSSM stereo matcher imposes a threshold of zero on the correlation score of potential matches before it accepts them. It must be concluded that this threshold was sufficient to suppress the MSSM stereo matcher from detecting any erroneous false matches.

Ideally, it would be expected that the results for Stereogram 4 (one noise image, `noise1.hips`, matched against another, `noise2.hips`) would be similar to those obtained for Stereogram 1 and Stereogram 3. However it can be seen from the output disparity map that a large number of random matches have been detected incorrectly by the MSSM stereo matcher. However it is also clear that these matches are of a consistently low confidence: the mean confidence was found to be 0.046 with a standard deviation of ±0.096. The most plausible explanation for this result is that it is again due to the finite size of the reference and search windows and the sub-optimal nature of the input Gaussian noise signal. This will inevitably lead to slight similarities occurring between small patches of the left and right image and hence result in random low confidence matches. A simple method of removing these

erroneous matches might be to increase the confidence threshold below which potential matches are rejected.

In summary, the performance of the MSSM stereo matcher does not vary significantly from that expected for the four standard examples of stereograms used as input. The slight variations from the ideal that were observed in the algorithm's performance can be attributed largely to the sub-optimal nature of the input noise signals. The band-pass filtering and windowed correlation calculations performed by the MSSM stereo matcher may also have adversely affect the algorithm's performance on these synthesised input images. However it is these very features that contribute greatly to MSSM stereo matchers success at matching real imagery. Finally, the results of this section provide strong evidence that MSSM algorithm has been implemented without serious errors and that the software performs without major bugs, except for the bug that was discovered in the calculation of the inter-scale confidence weights.

9.4. Accuracy versus integer disparity

9.4.1. Objective

The objective of this experiment was to investigate whether there was any correlation between the matching accuracy achieved by the MSSM stereo matcher and integer disparities, i.e. for an offset of an exact multiple of one pixel between the left and right images.

9.4.2. Method

Uniform disparity maps of size 129 × 129 pixels and with integer disparities from −16.0 to +16.0 pixels were synthesised using the HIPS `genframe` command. The right and left random noise images comprising the input stereograms were then generated using the method that was described in section 9.2 (although this was not strictly necessary since only stereograms containing full pixel disparities were required). After matching each input stereogram using the MSSM stereo matcher, the central 65 × 65 pixel region of the output disparity map was extracted to avoid any possibility of edge effects influencing the results. The error between this region of the output disparity map and the original synthetic disparity map was then recorded. The experiment was repeated for 50 stereograms generated with different random number seeds for each trial disparity.

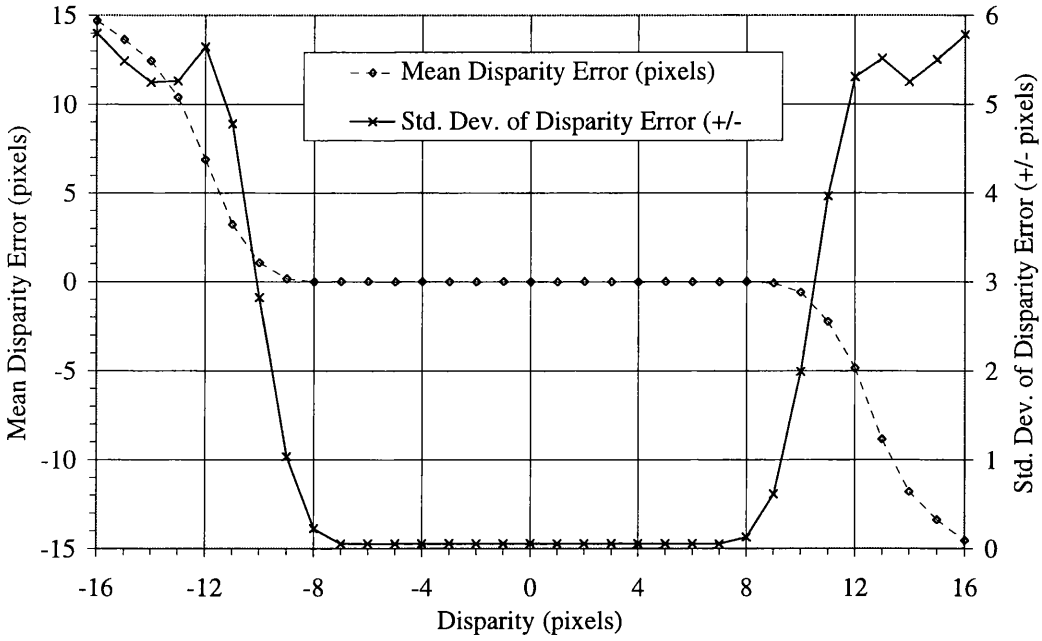


Figure 9.5. Graph of mean disparity error (pixels) and standard deviation of disparity error (pixels) against integer disparity (pixels).

9.4.3. Results

The mean error and standard deviation of the errors between the output disparity maps and the original synthesised disparity map have been plotted against integer disparity in figure 9.5. The mean disparity error was observed to be within ±0.010 pixels for integer disparities within the range –7.0 pixels to +7.0 pixels. The standard deviation of the disparity error is also very low within this range of integer disparities, being constant at approximately ±0.056 pixels to within ±0.002 pixels. However the disparity error starts to rise sharply when the disparity increases above approximately +8.0 pixels and decreases below approximately –8.0 pixels.

9.4.4. Interpretation of results

The most probable reason for the observed increase in the disparity error for large absolute disparities is that ±8.0 pixels is the limit of the MSSM stereo matcher's search offset for the input image size of 129 × 129 pixels. The stereo matcher will therefore fail to correctly match pixels with a disparity above 8.0 pixels or below –8.0 pixels.

However, it is likely that the correlation function for pixels with disparities of +8.0 pixels and above will still rise towards positive disparities and consequently there will

be a positive bias to the output disparity values. It is similarly likely that a negative bias will occur for pixels with disparities of –8.0 pixels or below. Therefore the increase in disparity error that is observed in figure 9.5 does not rise quite as sharply as would be expected if stereo matching simply failed completely for disparities beyond the matching limit.

The observed performance of the MSSM stereo matcher is excellent within the matching limit with an RMS disparity error of less than ±0.06 pixels. It is also noteworthy that the performance of the matcher is relatively constant within this disparity range with no observed bias to any particular value of disparity.

9.5. Accuracy versus sub-pixel disparity

9.5.1. Objective

The objective of this experiment was similar to that of the previous experiment with the exception that the performance of the MSSM stereo matcher was to be investigated for a range of sub-pixel disparities.

9.5.2. Method

Exactly the same method was employed as for the previous experiment with the exception that a range of disparity maps were synthesised with a uniform disparity of from –1.00 to +1.00 pixels in steps of 0.02 pixels. 50 random noise stereograms with different random number seeds were again generated from each uniform disparity map and matched using the MSSM stereo matcher.

9.5.3. Results

The standard deviation of the measured disparity error has been plotted against sub-pixel disparity in figure 9.6. The mean disparity has not been plotted because it was deemed to be insignificant (within ±0.010 pixels) for all disparities. It can be seen from figure 9.6 that the standard deviation of the disparity error dips from approximately ±0.056 pixels for a disparity of ±1.00 pixels to a minimum of approximately ±0.053 pixels between –0.50 and –0.40 pixels and between +0.40 and +0.50 pixels, rising to a maximum of approximately 0.058 pixels in the vicinity of 0.00 pixels disparity.

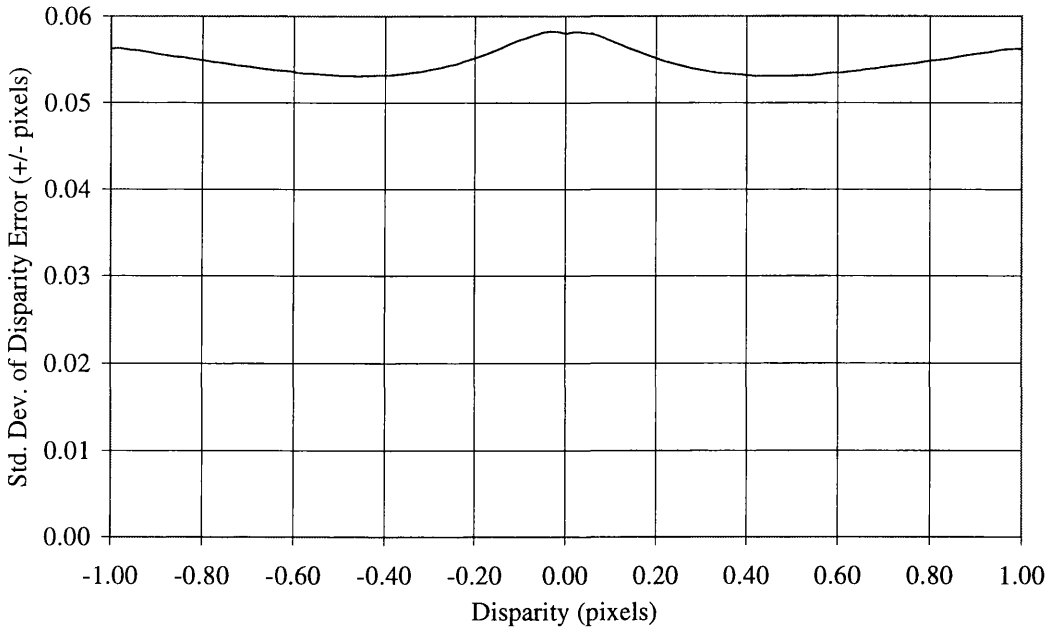


Figure 9.6. Standard deviation of disparity error (pixels) plotted against sub-pixel disparity (pixels).

9.5.4. Interpretation of results

Contrary to expectations, and fears, prior to the experiment, the MSSM stereo matcher actually appears to perform marginally better for sub-pixel disparities than it does for integer disparities. However, the observed performance difference was so slight (±0.005 pixels between the minimum and maximum standard deviation of disparity error) that it was not deemed necessary to carry out an investigation into its cause.

9.6. Accuracy versus disparity gradient

9.6.1. Objective

The objective of this experiment was to investigate the relationship between the gradient (i.e. rate of change) of the disparity in a disparity map and the resultant accuracy of the MSSM stereo matcher.

9.6.2. Method

This experiment's requirement for disparity maps with a known disparity gradient was satisfied by synthesising disparity maps of size 129 × 129 pixels with a uniform disparity ramp and hence a uniform disparity gradient across the entire image. 101

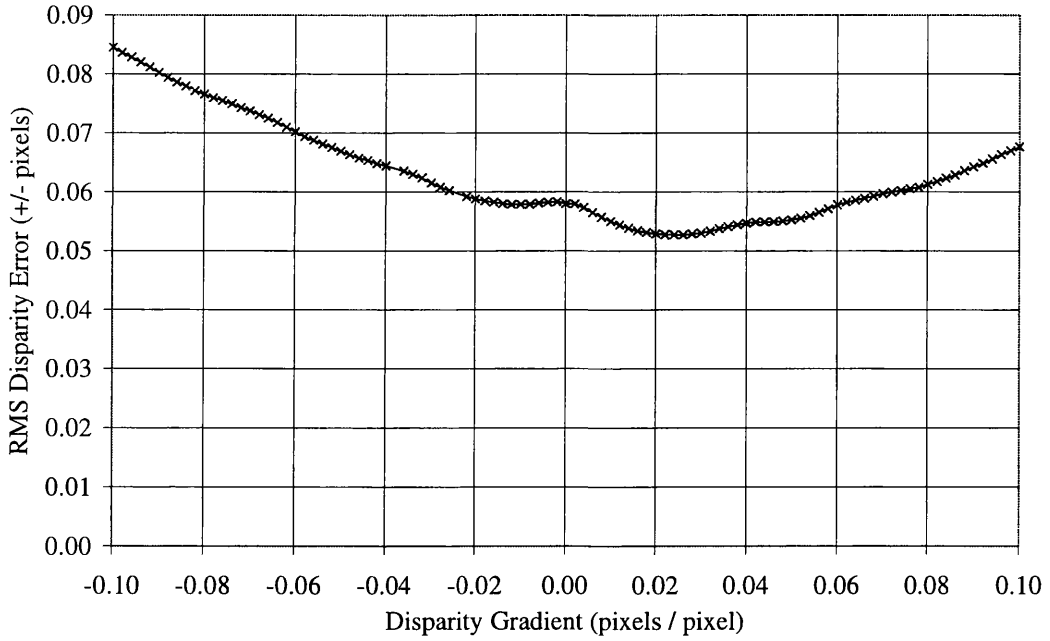


Figure 9.7. Graph of RMS disparity error (pixels) against disparity gradient (pixels / pixel).

such disparity maps were generated with disparity gradients ranging from –0.100 to +0.100 in steps of 0.002. A series of 50 random noise stereograms, generated with different random number seeds, was generated from each disparity map using the method described in section 9.2.

Each of the generated stereograms was then matched using the MSSM stereo matcher and the error between the obtained disparity map and the original synthesised disparity maps was computed for the central 65 × 65 pixel region of each. Extracting this central region from the output disparity maps ensured that a 32 pixel wide image border was maintained. This was especially important since the stretching/shrinking effect of disparity gradients that was described in sub-section 9.1.2 would tend to amplify image edge effects.

9.6.3. Results

The results for this experiment are shown in figure 9.7 which plots RMS disparity error measured in the central region of interest of the output disparity maps against the disparity gradient.

9.6.4. Interpretation of results

Figure 9.7 shows that the observed relationship between the disparity gradient and the disparity error is broadly in line with the expected result: namely that the disparity error increases with the magnitude of the disparity gradient. During the matching process, the MSSM stereo matching algorithm uses a single disparity value to interpolate the values of all pixels within the search window. Although it would be more accurate to generate the search window by interpolating with the current disparity estimates for all pixels within the window, this is not done because it would be extremely expensive computationally. Consequently the MSSM stereo matcher does not account for the relative stretching or shrinking of the right image caused by a disparity gradient. Therefore it is to be expected that the greater the disparity gradient between the left and right images, the more dissimilar will be the contents of the reference and search windows during the matching process, and the greater will be the disparity error.

Although the graph in figure 9.7 shows roughly the expected relationship of disparity error increasing with the magnitude of the disparity gradient, it seems to be slightly unusual on two counts. Firstly, it appears that the minimum disparity error does not occur at the expected position of zero disparity gradient, where there is no image stretching or shrinking and the left and right images are therefore identical, but at a disparity gradient of approximately +0.025. Secondly, disparity error appears to rise more sharply with positive disparity gradients than it does with negative disparity gradients. It was decided to perform another experiment to verify whether this somewhat unusual effect was actually resulting from the matching process. This experiment is described in section 9.7.

9.7. Verification of disparity gradient results

9.7.1. Objective

The results of the previous experiment seemed to indicate that there was a bias in the MSSM stereo matcher towards positive disparity gradients. However, an alternative explanation for this unusual characteristic might be that the observed bias actually lies in the image generation process rather than in the MSSM stereo matcher itself. That is, it was possible that the sub-pixel interpolation used to generate the left images was affecting their *monocular* characteristics in such a way that those generated for positive disparity gradients were somehow pre-disposed to match with more accuracy than those generated for negative disparity gradients. The objective of this experiment

was to investigate whether this was indeed the case, or whether the MSSM stereo matcher is actually biased towards positive disparity gradients.

9.7.2. Method

This experiment was performed by first generating a right Gaussian noise image, `right.hips`. Two left images, `left_neg.hips` and `left_pos.hips`, with disparity gradients of –0.05 and +0.05 respectively were then synthesised from `right.hips` using the method described for the previous experiment. `right.hips`, `left_neg.hips` and `left_pos.hips` were matched against themselves and the resulting RMS disparity error was recorded. These RMS disparity errors were compared with those obtained when `left_neg.hips` and `left_pos.hips` were matched against `right.hips` in the more usual way as performed in the previous experiment. The experiment was repeated 50 times, with a different random seed used to generate `right.hips` each time.

9.7.3. Results

The RMS disparity errors obtained for all of the random noise seeds for each of the five sets of matches described in sub-section 9.7.2 are presented in table 9.3 below.

Left image	Right image	RMS disparity error (pixels)
`right.hips`	`right.hips`	±0.058
`left_pos.hips`	`left_pos.hips`	±0.060
`left_neg.hips`	`left_neg.hips`	±0.061
`left_pos.hips`	`right.hips`	±0.055
`left_neg.hips`	`right.hips`	±0.068

Table 9.3. Comparison of the RMS disparity error obtained when left images generated for disparity gradients of –0.05 (`left_neg.hips`) and +0.05 (`left_pos.hips`) are matched against themselves and against the original right image (`right.hips`).

9.7.4. Interpretation of results

Table 9.3 shows that approximately the same RMS disparity error results when either `left_pos.hips` or `left_neg.hips` is matched with an exact replica of itself (i.e. forming stereograms with uniformly zero disparity). In both cases, the resulting RMS disparity error is slightly higher than that observed when the right image, `right.hips`, is matched with an exact replica of itself. Both `left_pos.hips` and `left_neg.hips` images are generated by sub-pixel interpolation from the right image, while the right image is not subject to this interpolation process. This implies that the process of sub-pixel interpolation reduces slightly the inherent "matchability" of either image regardless of their disparity gradient. This is probably due to the low-pass filtering effect of the pixel interpolation process. In fact, it should be noted that the RMS disparity error observed for the "positive disparity gradient" image, `left_pos.hips`, is actually very slightly higher than that observed for the "negative disparity gradient" image, `left_neg.hips`.)

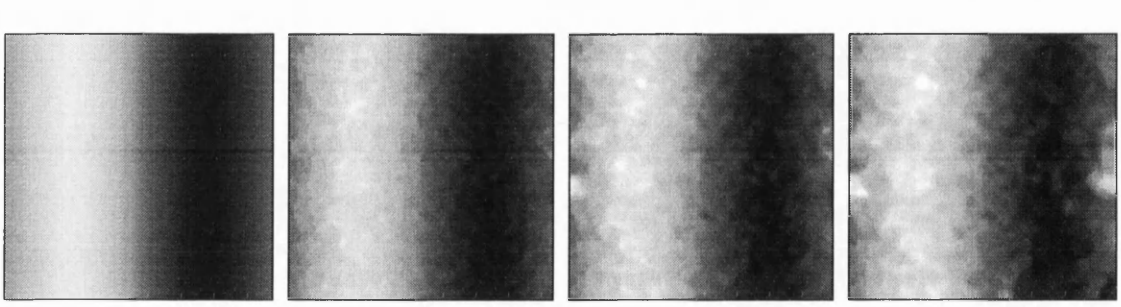
Table 9.3 also shows that RMS disparity error obtained from matching the +0.05 disparity gradient left image, `left_pos.hips`, with the original right image, `right.hips`, is slightly lower than that obtained from matching either `left_pos.hips` or `right.hips` with an exact replica of itself. Conversely, the RMS disparity error obtained when the –0.05 disparity gradient left image, `left_neg.hips`, is matched with the original right image, `right.hips`, is slightly higher than when either is matched with an exact replica of itself.

These results imply that the bias towards positive disparity gradients observed in the previous experiment is not due to the sub-pixel interpolation process altering the intrinsic monocular characteristics of the images. It can therefore be concluded that the MSSM stereo matcher does indeed react differently to positive and negative disparity gradients. However, the results of this experiment do not help to explain the reason for this bias. This question therefore remains a suitable topic for future investigation.

9.8. Accuracy versus image noise

9.8.1. Objective

The objective of this experiment was to characterise the degradation in the accuracy of the MSSM stereo matcher with increasing quantities of image noise.



Input disparity *S/N = 10.0dB* *S/N = 5.0dB* *S/N = 2.5dB*

Figure 9.8. The synthetic input disparity map and example output disparity maps for the signal-to-noise-ratios of 10.0dB, 5.0dB and 2.5dB. All are shown scaled from −5.0 (black) to +5.0 (white).

9.8.2. Method

A 128 × 128 pixel disparity map consisting of a horizontal sinusoid of period 128 pixels and amplitude 4.0 pixels was synthesised for use as the input to this experiment in order to provide a wide range of disparity gradients. The left and right random noise images, with a noise standard deviation, σ_s, of 32 grey-levels, were generated from this synthesised disparity map using the method described in section 9.2. Gaussian noise, with standard deviation σ_n, was then added to the left and right images. The three random number seeds used to generate the Gaussian noise added to the left image, added to the right image, and used to generate the original stereogram were all different. 50 such random noise stereograms were generated, all with different random number seeds, and matched using the MSSM stereo matcher. The experiment was repeated with σ_n ranging from 0 (i.e. no additional noise) to 32 (added noise equal to the original signal).

9.8.3. Results

The input synthesised disparity map and example output disparity maps for signal-to-noise-ratios of 10.0dB, 5.0dB and 2.5dB are shown in figure 9.8. Figure 9.9 shows plots of slices along row 64 of these disparity images. For each noise standard deviation, σ_n, the signal-to-noise-ratio was calculated using the following equation:

$$S/N\,(dB) = 20\log_{10}\left(\frac{\sigma_s}{\sigma_n}\right) \tag{9.4}$$

The RMS error in the output disparity maps is plotted against the signal-to-noise-ratio of the input stereograms in figure 9.10.

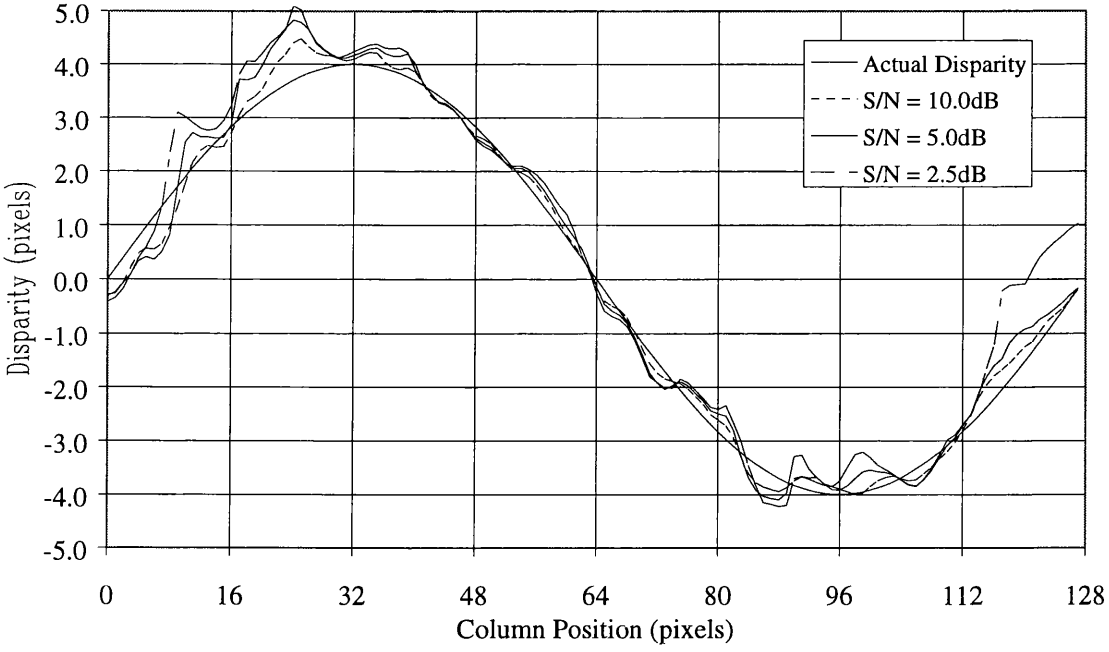


Figure 9.9. Plot of slices along row 64 of the disparity images shown in figure 9.8.

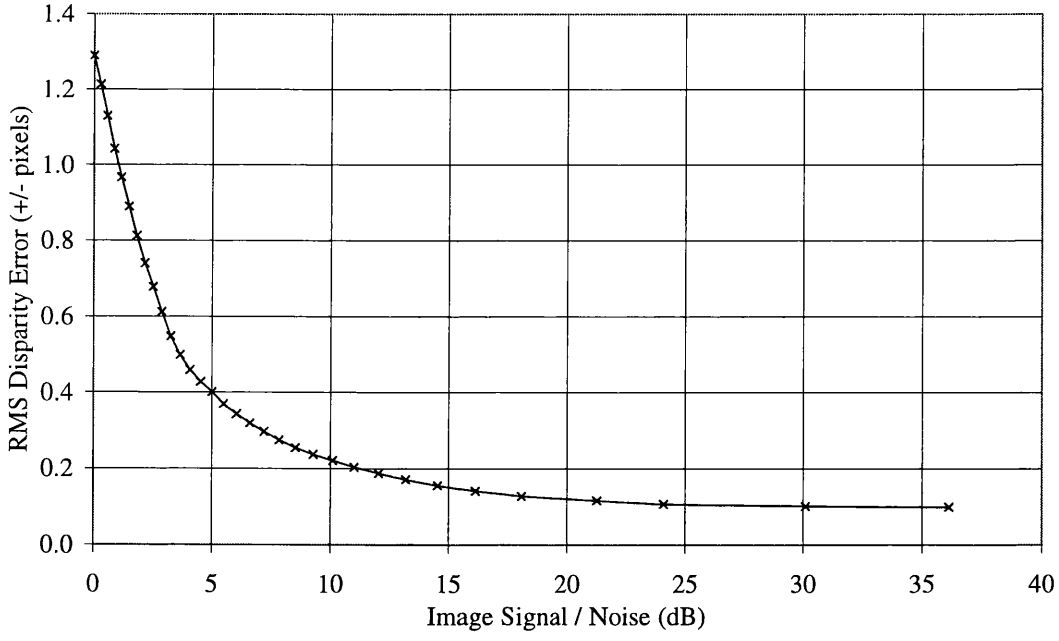


Figure 9.10. Graph of RMS disparity error (pixels) against image signal-to-noise-ratio (dB).

9.8.4. Interpretation of results

It can be seen from figure 9.10 that the accuracy of the MSSM stereo matcher is approximately ±0.1 pixels RMSE for signal-to-noise-ratios better than 25dB and that the accuracy degrades gracefully for signal-to-noise-ratios worse than 25dB.

9.9. Conclusions

During the investigations described in this chapter, the highest accuracy achieved by the MSSM stereo matcher was observed to be ±0.053 pixels RMSE. This figure was obtained for an input stereogram comprising Gaussian random noise images, a uniform input disparity map and no image noise. The somewhat idealised nature of this input stereogram means that it is highly unlikely that a similarly high level of accuracy will be achieved when the MSSM stereo matcher operates as part of the ASP stereo vision system. However, a match accuracy of ±0.053 pixels RMSE forms a convenient benchmark for gauging the relative performance of the algorithm when applied to different stereograms.

The accuracy of the stereo matcher was observed not to vary with integer disparities within its matching limit and to vary only very slightly (by approximately ±0.003 pixels RMSE) with sub-pixel disparities. Surprisingly, the best accuracy was observed to occur for a uniform disparity value of approximately ±0.45 pixels and not for a disparity of zero as was initially expected. Further investigations will be required if the reason for this slight bias towards sub-pixel disparities is to be ascertained. However, for all practical purposes, the accuracy of the MSSM stereo matcher can be assumed to be insensitive to either integer or sub-pixel disparities when operating within its matching limits. (In separate trials that have not been reported here, the disparity limit for full CCIR images was found to be approximately ±50 pixels.)

A graph of RMS disparity error against the signal-to-noise ratio of input stereograms was obtained in figure 9.10. This graph gives a good indication of the relative sensitivity of the MSSM stereo matcher to different levels of image noise. For the low levels of image noise produced by current high quality CCD cameras, the expected RMS disparity error is approximately ±0.1 to ±0.2 pixels RMSE. By contrast to the input stereogram used to obtain the benchmark accuracy figure of ±0.053 pixels RMSE, these results were obtained for a more representative type of input disparity map. Therefore they are perhaps a more reasonable estimate for the order of accuracy to be expected from the MSSM stereo matchers when it is used within the ASP stereo vision system.

However, it should be noted that the experiments described in sections 9.6 and 9.7 showed that the accuracy of the MSSM stereo matcher is fairly dependant on the shape (e.g. disparity gradient) of the underlying disparity map as well as the signal-to-noise ratio of the stereo images. Therefore, because the graph of figure 9.10 was obtained for only one particular sinusoidal disparity map, it can not be used to determine precisely the accuracy with which an arbitrary stereogram will be matched purely from an estimate of its signal-to-noise ratio.

Nevertheless a figure of between ±0.1 to ±0.2 pixels RMS disparity error seems to be a reasonable estimate for the expected global accuracy of the MSSM stereo matcher. If it is assumed that the actual match accuracy is ±0.2 pixels RMSE and that the full disparity dynamic range of 100 pixels (i.e. ±50 pixels) is utilised, then a rough estimate for the average expected depth accuracy would be $\pm^{1}/_{500}$th of the viewing volume depth. For the viewing volume depth of 500mm specified for the ASP stereo vision system this would equate to an approximate average depth measurement accuracy of ±1.0mm RMSE. This is well within the stereo matcher requirements specified for the ASP stereo vision system.

In addition to the information gained about the performance of the MSSM stereo matcher, this analysis has also resulted in the development of a series of performance tests in the form of UNIX shell scripts. These could easily be adapted to compare the relative accuracies of other stereo matchers. The work described in this chapter has also provided a great deal of understanding about the mechanisms by which stereo images are formed. This has resulted in the creation of many opportunities for further work in this area. Finally, it should be noted that, to date, no other such accuracy analysis of a stereo matcher has been reported in the computer vision literature.

Chapter 10
Static Calibration

10.1. Introduction

10.1.1. Background

It was established early in this thesis that two problems must be addressed by a videometrics system such as the Active Stereo Probe that is intended to obtain quantitative 3D measurements of an observed scene. The first of these is to determine automatically and accurately the point-wise correspondences between the left and right images in a stereogram of the scene. The successful development and performance analysis of a suitable stereo matcher was described in the two previous chapters. Having thus established the correspondence between left and right image points, a videometrics system must then address the second problem of *space intersection* to triangulate the three-dimensional position of the original world point. However this requires that the *internal orientation parameters* (the focal length and image centre and possibly additional lens distortion parameters) and the *external orientation parameters* (the position and orientation of the cameras) of the system be known. These parameters are determined by a process of *calibration.*[1]

Chapter 3 describes the two stage approach to calibration that was proposed for use in the ASP stereo vision system. The first stage of *static calibration* derives estimates for both the internal and external orientation parameters while the cameras are held in a fixed position. Thereafter, the second stage of *dynamic calibration* maintains an updated estimate of the external orientation parameters when the cameras move. This chapter describes the development of the static calibration module for the ASP stereo vision system.

[1] In photogrammetric terminology, *camera calibration* is performed to determine the internal orientation parameters of a camera while *space resection* is performed to determine the external orientation parameters.

As the ASP project proceeded, it became increasingly obvious that many applications could be satisfied by using the ASP stereo vision system in a static mode of operation in which the cameras are held in fixed positions. In fact, a new project was started to develop a prototype product called CopyCAD that would use the basic ASP stereo vision system in a static mode in one such application area: maxillo-facial measurement. It was therefore decided to concentrate the development of the static calibration module to address the immediate requirements of this application rather than the somewhat less stringent requirements originally specified for the ASP stereo vision system. As a consequence, the problem of maintaining the dynamic calibration of a mobile camera system was no longer within the scope of the current project. However a dynamic calibration scheme employing the proposed method of actuator feedback is presented in the further work described in Chapter 12.

10.1.2. Objective

The primary objective for the development of the static calibration module was that it should provide the means both to calibrate the ASP stereo vision when operating statically and subsequently to perform space intersection to recover 3D surfaces from input disparity maps. It was intended that the developed static calibration module should be sufficiently accurate to meet the requirements of maxillo-facial measurement. These were essentially that surface recovery should be carried out with a minimum accuracy of ±1.0mm RMSE over the volume of a human face. (In later work the required accuracy was revised to ±0.5mm RMSE after extensive consultation with maxillo-facial clinicians.)

In addition to this accuracy requirement, it was also deemed necessary that the static calibration module should be as simple and easy to use as possible so that static calibration could be carried out "in the field" by non-specialist users. Although the time taken to perform static calibration was not required to be as short as that essential to dynamic calibration, it was still required to remain in keeping with the on-line, interactive nature of the ASP stereo vision system.

10.1.3. Approach

The standard method used for the calibration of fixed camera 3D vision systems is to image a measured grid of target points and to detect the position of the resulting image points. Various techniques have been developed that then allow the external and internal orientation parameters to be derived from the known 3D positions of the target points and the detected positions of their corresponding image points. This

approach to calibration was deemed to be perfectly acceptable and was therefore adopted as the basis of the static calibration module.

Three main problems are therefore required to be addressed by the static calibration module of the ASP stereo vision system:

1. The provision of some form of calibration grid that comprises targets whose 3D world co-ordinate positions are accurately known.

2. The development of a method to detect automatically the image location of the target points.

3. The implementation of a method to determine the internal and external orientation parameters from the sets of known world co-ordinates and detected image co-ordinates.

It is clear that the first two problems are intimately related since the characteristics of the target points will determine how easily and accurately their locations can be detected in resulting images. In [Cumani, 1991] it is shown that the image location of the centres of circular discs can be detected with extremely high accuracy. Circular discs were therefore chosen as the target points to be used in the static calibration module of the ASP stereo vision system. Consequently, a calibration grid was constructed to provide an appropriate array of circular targets as described in section 10.2. Section 0 details the development of an appropriate target detector based upon the one described in [Cumani, 1991].

The many methods that have been investigated to address the third of the above problems are summarised in Chapter 2. One such method, *the direct linear transform* (DLT) [Abdel-Aziz, 1971], was chosen for investigation. Although it was originally developed for conventional photographic cameras, the DLT has been shown to be equally effective for calibrating CCD cameras [Burner, 1985]. It was chosen for initial investigation because it was fast and simple to implement and hence allowed a basic static calibration module to be implemented quickly. The software developed for calibration and subsequent space intersection using the DLT and some results obtained using it are given in section 10.4.

10.2. Design and construction of the calibration grid

10.2.1. Requirements

The main requirements of the calibration grid were:

1. There should be a strong contrast between the target circles and the background. so that they can be detected easily and accurately in images of the calibration grid.

2. The three-dimensional positions of the centres of the target circles should be known with sufficient accuracy.

3. The target points should be distributed as evenly as possible throughout the required viewing volume. Otherwise it is likely that calibration will be more accurate in regions that are dense in target points and less accurate in regions that are more sparsely populated with target points.

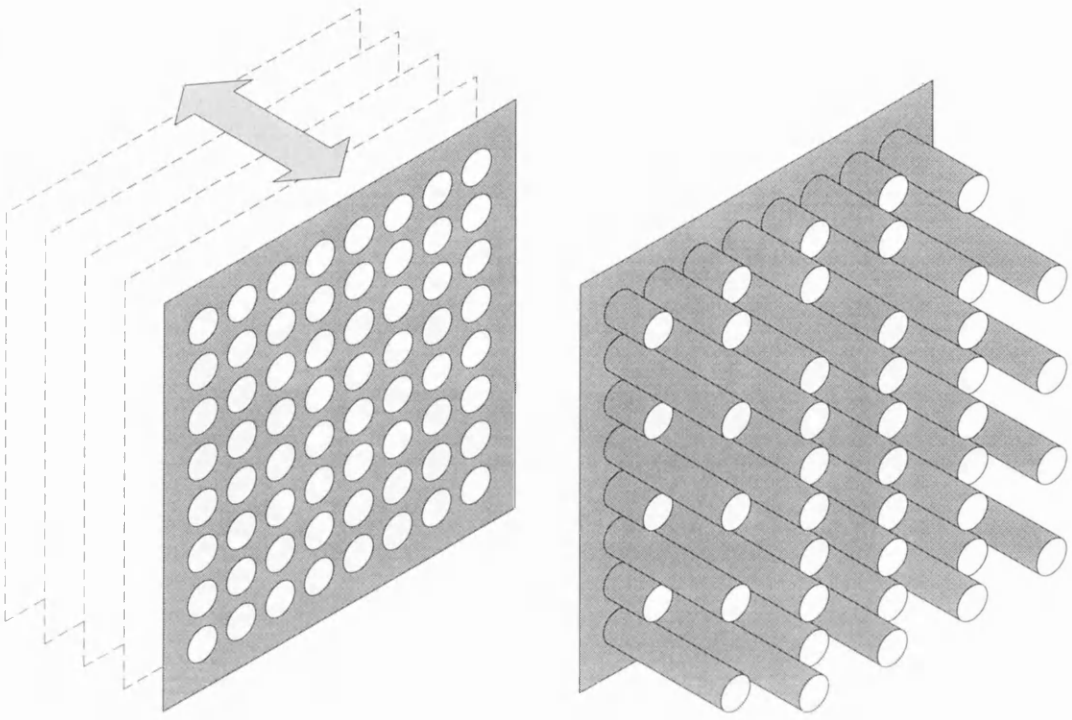
4. The targets should be as close to perfect circular discs as possible. This is important because exact circularity of the targets is assumed when the centres of the targets are detected.

5. None of the target points should be occluded in images of the calibration grid. If some targets are partially occluded by others, it makes their detection much more difficult and less accurate than for wholly visible targets.

10.2.2. Possible configurations

The two basic configurations that are shown in figure 10.1 were considered for use as the target grid. The first is a plane of contrasting circular targets that are evenly distributed over the X and Y dimensions of the working volume. A distribution of target points over the Z dimension of the working volume is obtained by translating the plane backwards and forwards. The other configuration that was considered is an array of parallel cylinders of different lengths. The circular ends contrast with the sides of the cylinders and the base plane forming the grid of target discs. The different positions and lengths of the cylinders ensures that an even distribution of targets over the working volume is obtained.

The translating plane approach is the one that has been most widely adopted in the literature. It offers the advantages that a plane of highly accurate contrasting circles can be obtained quite easily by photographic processes and that a larger number of target points can be obtained by translating the plane to different positions. However



Translating plane of circular targets. *Fixed array of cylinders.*

Figure 10.1. Two basic configurations considered for the calibration grid.

the requirement to physically translate the plane results in a number of disadvantages to this approach. Firstly a mechanical device, such as a stepper motor translation stage, has to be used which increases both the cost and complexity of the calibration grid itself. Secondly, the requirement to move the plane to a number of known positions, capturing images at each, significantly increases the complexity of the calibration procedures. Finally, it was felt that the use of a translation stage would result in less accuracy than expected due to the inevitable inaccuracy in positioning the stage itself and the requirement that it be positioned exactly perpendicular to the target plane.

By comparison with the translating plane approach, the fixed array of cylinders is perhaps more difficult to construct accurately and provides fewer targets of the same size. However a calibration procedure based upon a fixed array of cylinders will be much less complicated since only one image of a static grid is required. In addition, a calibration grid comprising a fixed array of cylinders was felt to be much less likely to go out of calibration itself because it has no moving parts. The fixed array of cylinders approach was therefore chosen because the fundamental simplicity of the calibration procedures required by it were felt to far outweigh the slight advantages of the translating plane approach.

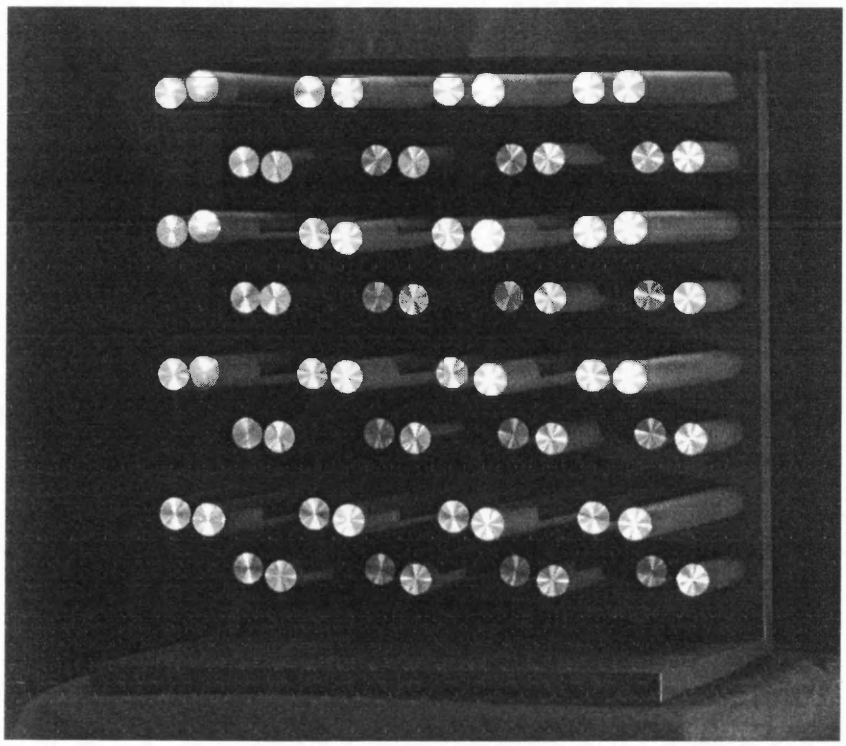


Figure 10.2. The prototype calibration grid.

10.2.3. Prototype calibration grid

Number of cylinders	64 (8 × 8)
Horizontal spacing	40mm
Vertical spacing	40mm
Cylinder diameter	18mm
Cylinder lengths	16 cylinders × 20mm 16 cylinders × 80mm 16 cylinders × 140mm 16 cylinders × 200mm

Table 10.1. Basic dimensions of the prototype calibration grid.

Figure 10.1 shows a prototype calibration grid that was kindly constructed by technicians at the Department of Manufacturing Engineering and Management at the University of Strathclyde. Its can be seen from the basic dimensions of the prototype calibration grid that are given in table 10.1 above that it occupies a volume of

approximately 300mm × 300mm × 180mm. It was originally intended that this small prototype design would be fully tested before commencing the construction of a full sized one meeting the viewing volume requirements of the ASP stereo vision system that were given in Chapter 3.

Before the prototype calibration grid was constructed, it was first simulated in the ASP sensor system model that was described in Chapter 4. A number of experiments were thus carried out to ensure that none of the target circles would be occluded in either of the stereo images captured for any position of the calibration grid within the specified working volume of the ASP stereo vision system.

After a suitable design had been drawn up, it was passed to technicians at the Department of Manufacturing Engineering and Management at the University of Strathclyde for manufacture. The cylinders were each machined from solid aluminium rod and pressed into an aluminium base-plate. The whole structure except for the circular ends of the cylinders was then painted matt-black by spraying it while the ends were covered in masking tape. It was later found that too much specular reflection resulted from the bare aluminium of the circular targets, so these were spray-painted in matt grey. Great care was taken to ensure that a sharp change in contrast was maintained between the bodies of the cylinders and their ends.

It was originally intended that the manufacturing process would be sufficiently accurate that the nominal positions of the cylinders could be taken from the drawings and used as the target locations for calibration. However, due to a fault in the manufacturing process, the cylinders were not fitted to the base plate correctly and it was therefore necessary to measure the positions of the cylinder ends. This was done to an accuracy of approximately ±10μm using a micrometer. Unfortunately, it was later found that some of the cylinders were loose, clearly meaning that their measured positions would be subject to significant amounts of error. The least mechanically stable of the calibration cylinders were removed and the prototype calibration grid was used successfully to verify the operation of the static calibration module.

10.2.4. Replacement calibration grid

Eventually the lack of accuracy of the prototype calibration grid forced it to be abandoned and the replacement calibration grid that is shown in figure 10.3 was designed and constructed. The replacement calibration grid was specifically intended to be used for use in the maxillo-facial measurement application and was hence designed to fill the approximate volume of a human head as shown in table 10.2.

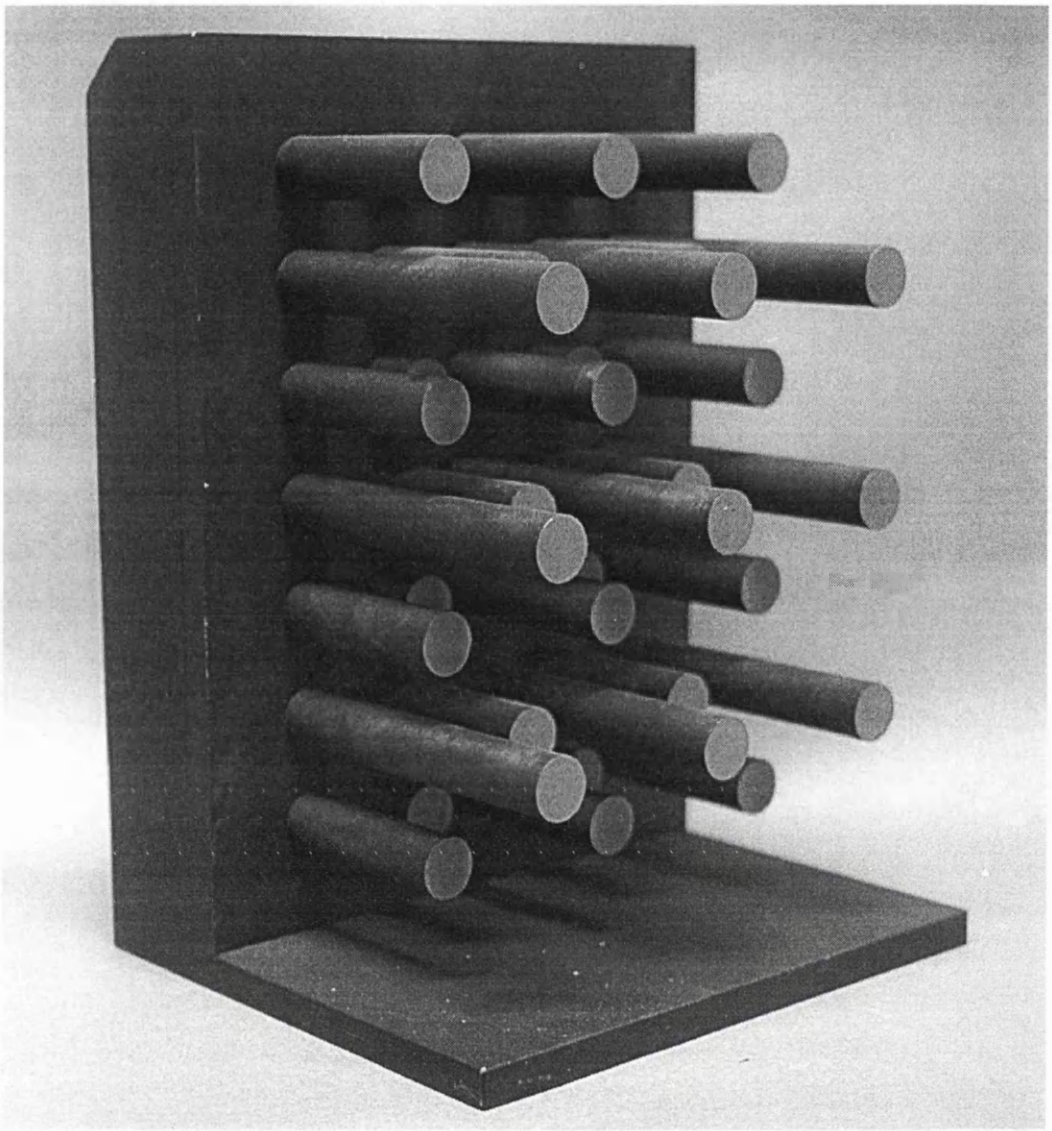


Figure 10.3. The replacement calibration grid.

Number of cylinders	35 (7 × 5)
Horizontal spacing	35mm
Vertical spacing	35mm
Cylinder diameter	20mm
Cylinder lengths	8 cylinders × 20mm 9 cylinders × 70mm 18 cylinders × 100mm

Table 10.2. Basic dimensions of the replacement calibration grid.

The same design and manufacturing process as that of the prototype was used for the replacement calibration grid except that it was successfully manufactured by a specialist sub-contractor to tolerances of ±0.002" (approximately ±5μm). The nominal positions of the targets have not yet been verified by measurement.

10.3. Target detector development

10.3.1. Requirements

The basic requirements of the target detector were that it should be able to detect automatically, robustly and with high accuracy the location of the centre of the target circles in images of the calibration grid. An additional, although less rigorous, requirement was that the target detector should also be capable of processing a full 576×768 pixel image in a time commensurate with the interactive nature of the ASP stereo vision system. A final requirement of the target detector was that it must be possible to associate the detected image locations of the target circles with the corresponding 3D world locations.

10.3.2. Contour tracing

The first stage of the method of detecting circular targets that is described in [Cumani, 1991] is to find the bounding contour of all of the imaged targets. The method used is based upon that originally detailed in [Grattoni, 1990] for tracing the contour of contrast changes in an image to sub-pixel accuracy. This proceeds by first filtering the input image with a difference of Gaussians (DOG) function. (In [Grattoni, 1990] it is suggested that the second directional derivative of Gaussian filtered images be used. However, it can be seen from the results presented in section 10.3.4 that the use of an isotropic DOG filter still provides highly accurate results.)

The zero crossings in the DOG filtered (floating point) image correspond to strong contrast changes in the original image. The next task is therefore to detect the location of the zero-crossings and hence contrast boundaries to sub-pixel accuracy. A zero crossing occurs between a pixel, (x_i, y_i) and one of its neighbours, (x_j, y_j), if their pixel values have opposite signs. i.e. if the values of the two pixels are $F(x_i, y_i)$ and $F(x_j, y_j)$ respectively then a zero crossing occurs in the space between them if:

$$F(x_i, y_i) \cdot F(x_j, y_j) < 0 \qquad (10.1)$$

The implementation of the contour detector proceeds using the four steps that are illustrated in figure 10.4 and described below:

Step 1: The image is scanned one pixel at a time in column and row order until a zero crossing is detected. The presence of a zero crossing at a pixel location, (x, y), is determined by comparing the values of the four pixels at locations (x, y), $(x+1, y)$, $(x, y+1)$, and $(x+1, y+1)$:

$$a_{00} = F(x, y) \tag{10.2}$$

$$a_{10} = F(x+1, y) \tag{10.3}$$

$$a_{01} = F(x, y+1) \tag{10.4}$$

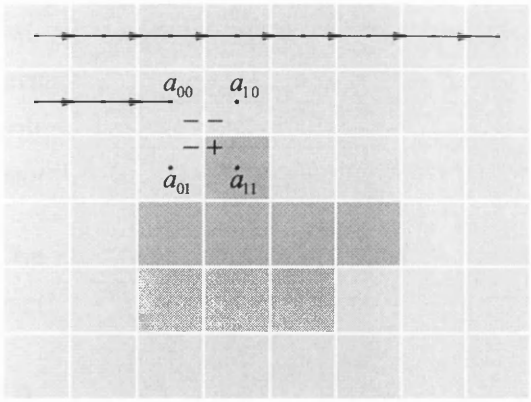
$$a_{11} = F(x+1, y+1) \tag{10.5}$$

Inequality	*Side*	*Z.C. Location*	α	*Next Pixel*
$a_{00} \cdot a_{10} < 0$	Top	$(x+\alpha, y)$	$\frac{a_{00}}{a_{00} - a_{10}}$	$(x, y-1)$
$a_{10} \cdot a_{11} < 0$	Right	$(x+1, y+\alpha)$	$\frac{a_{10}}{a_{10} - a_{11}}$	$(x+1, y)$
$a_{11} \cdot a_{01} < 0$	Bottom	$(x+\alpha, y+1)$	$\frac{a_{11}}{a_{11} - a_{01}}$	$(x, y+1)$
$a_{01} \cdot a_{00} < 0$	Left	$(x, y+\alpha)$	$\frac{a_{01}}{a_{01} - a_{00}}$	$(x-1, y)$

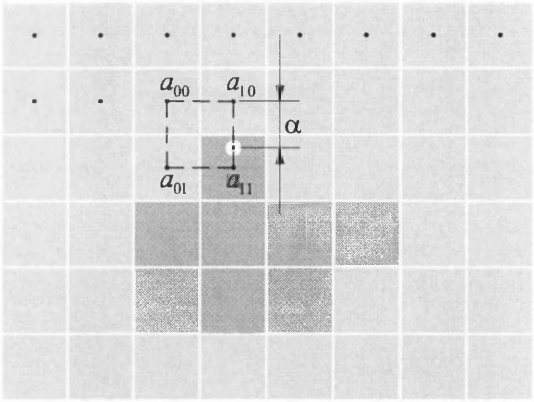
Table 10.3. Determination of the presence and location of a zero crossing contour for an image pixel location (x, y).

A zero crossing is present within the square bounded by the four points if any of the inequalities given in the first column of table 10.3 are true.

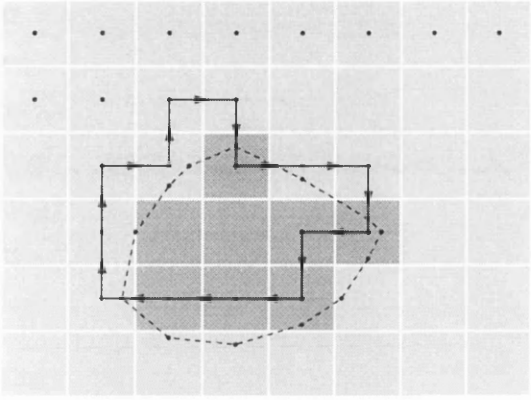
Step 2. Once a zero crossing is detected, the sides of the square bounded by the four points are checked to determine which intersect with the zero crossing contour as shown in the second column of table 10.3. The sides are checked in strict clockwise order: top first, then right, then bottom, then left. The location of the point of intersection that is found first becomes the first in a linked list of points that will eventually define the contour. The co-ordinates of the point of intersection have one integer and one sub-pixel co-ordinate as shown in the third column of table 10.3. The sub-pixel component of the co-ordinate, α, is calculated using the appropriate bilinear interpolation formula given in the fourth column of the table.



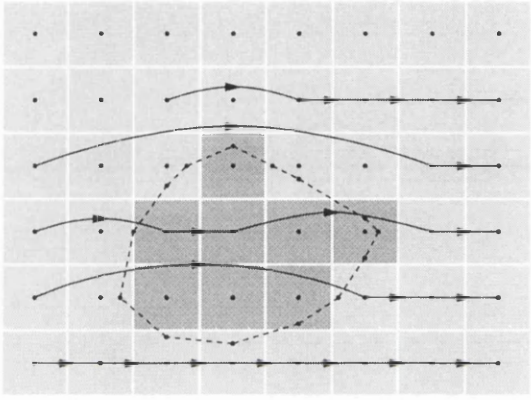
Step 1. The image is scanned until a zero crossing is detected.



Step 2. The location of the zero crossing is recorded.



Step 3. The contour is traced until it returns to the starting pixel.



Step 4. The contour is closed and normal image scanning is resumed.

Figure 10.4. Illustration of the contour tracing process: the light and dark squares represent pixels with values that have opposite signs.

Step 3. After a zero-crossing has been detected, normal scanning of the image is suspended until the zero-crossing contour has been completely traced. This is done by repeating step 2 for the next pixel on the contour whose location is shown in the last column of table 10.3. For the example illustrated in figure 10.4, the first point of intersection was found to lie on the right side of the square, i.e. between $(x+1, y)$, and $(x+1, y+1)$, so that the next pixel to be processed is the one to the right of (x, y), i.e. $(x+1, y)$. Step 2 is then repeated with the new pixel at the top left corner of the square and another point is added to the linked list of points and so on for subsequent pixels on the contour. Note that the search order for each pixel after the first is altered to ensure that the contour is traced out in a constant clockwise direction.

Step 4. The contour will be completely traced once processing returns to the pixel at which the zero crossing was first detected. The contour is then closed and the linked

list of sub-pixel co-ordinates defining the contour is returned. Note that contours that terminate at the edge of the image are ignored. Normal scanning of the image in column and row order is then resumed, skipping pixels that have already been processed during contour tracing.

The final result of contour tracing is therefore a list of M contours, $C_j, j = 1,2,\cdots,M$, each comprising a linked list of N_j points: $(x_{i,j}, y_{i,j}), i = 1,2,\cdots,N_j$.

10.3.3. Circle-finding

Static calibration requires the detection of the centres of the image contours belonging to the target circles, since it is the centres of the physical targets that are the known world target points. Unfortunately, the contour tracer was found to detect many contours, such as those due to the shadows of the calibration cylinders, in addition to those belonging to the images of the target circles. The circle-finding stage of the target detector therefore had not only to determine the position of the centres of the image contours but also had to determine which of the detected contours actually corresponded to the target circles.

The circle-finding stage proceeds by first estimating the location, (x_{cj}, y_{cj}), of the centre of every contour C_j using the method described in [Cumani, 1991]. This method treats the contour as a polygon of the N_j vectors between the detected points, $(x_{i,j}, y_{i,j})$. The centroid of the polygon is used as the estimate of its centre position:

$$x_{cj} = \sum_{i=1}^{N_j} \tfrac{1}{2} L_{i,j}\left(x_{i+1,j} + x_{i,j}\right) \Big/ \sum_{i=1}^{N_j} L_{i,j} \tag{10.6}$$

$$y_{cj} = \sum_{i=1}^{N_j} \tfrac{1}{2} L_{i,j}\left(y_{i+1,j} + y_{i,j}\right) \Big/ \sum_{i=1}^{N_j} L_{i,j} \tag{10.7}$$

Where $L_{i,j}$ is the length of the ith side of contour C_j:

$$L_{i,j} = \sqrt{\left(x_{i+1,j} - x_{i,j}\right)^2 + \left(y_{i+1,j} - y_{i,j}\right)^2} \tag{10.8}$$

After estimating the centre of the contour, the circle finder then calculates its mean radius, r_j, and an estimate of its circularity, Θ_j, which is the mean radius for the contour divided by the variance of the distance of each point from the centre of the contour:

$$r_j = \frac{1}{N_j} \sum_{i=1}^{N_j} r_{i,j} \tag{10.9}$$

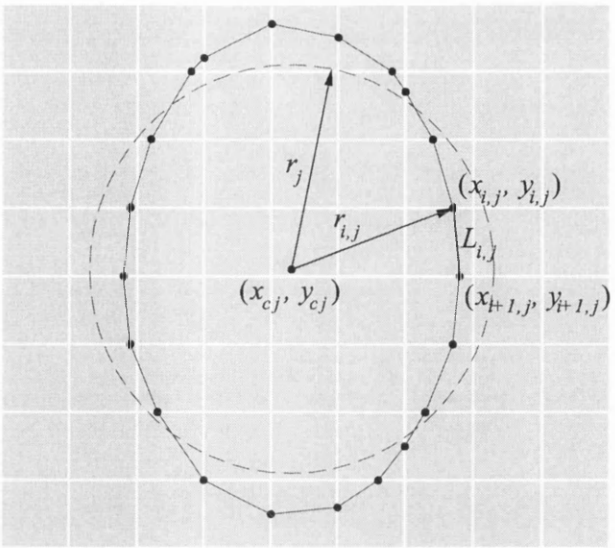


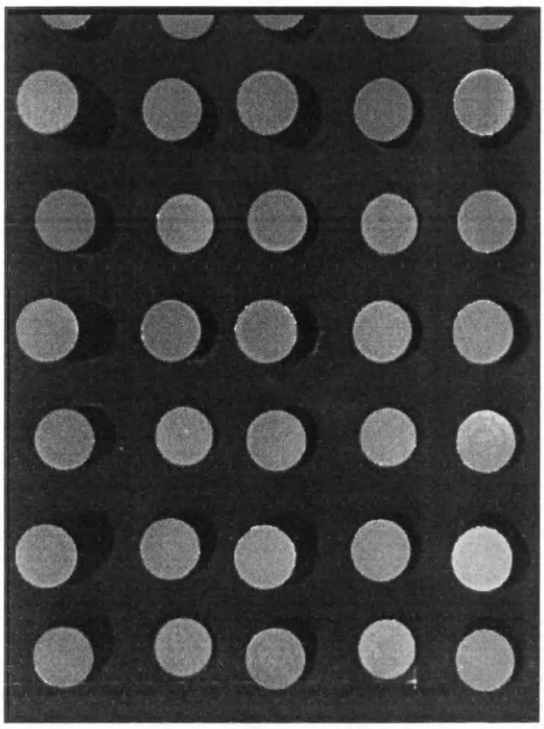
Figure 10.5. Illustration of the circle-finding process.

$$\Theta_j = \frac{r_j}{\frac{1}{N_j}\sum_{i=1}^{N_j}\left(r_{i,j} - r_j\right)^2} \tag{10.10}$$

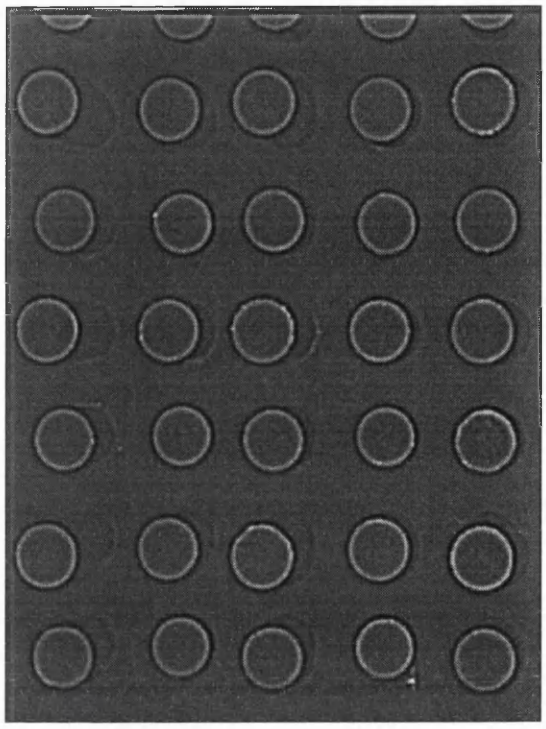
$$r_{i,j} = \sqrt{\left(x_{i,j} - x_{cj}\right)^2 + \left(y_{i,j} - y_{cj}\right)^2} \tag{10.11}$$

The assumption is made that contours of the target circles approximate to circles of a certain radius. Therefore a minimum mean radius and minimum circularity constraint was imposed to determine which of the detected contours actually corresponded to the target circles.

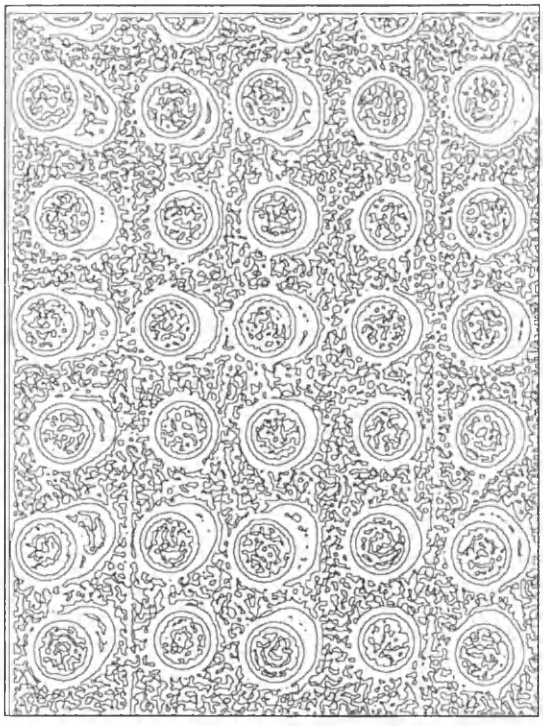
The complete target detection process is illustrated in figure 10.6..



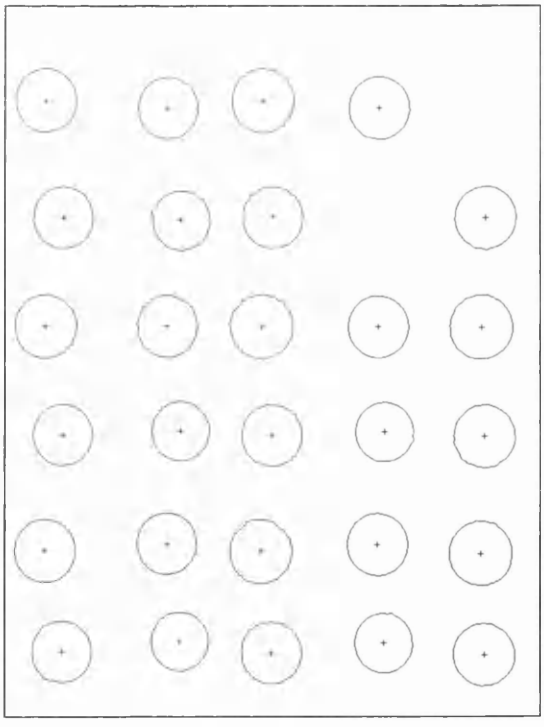
Typical input image.



DOG filtered image.



Zero crossings.



Detected targets.

Figure 10.6. Detection of circular targets in an image of the calibration grid.

10.3.4. Matching of world and image points

The final requirement of the target finder was that it must be possible to associate the detected image locations of the centres of the target circles with the corresponding 3D world locations. The fact that the target finder did not always find all of the visible targets somewhat complicated this task. Consequently, it was not possible to develop easily a fully automatic method for matching the world and image locations of the target points. Instead, a graphical user interface was developed as part of the CopyCAD system to allow the user to match manually the detected image and the known world locations of the targets. This procedure is described in more detail in Chapter 11.

10.3.5. Accuracy analysis

10.3.5.1. Objective

The aim of this experiment was to analyse the accuracy of the developed target detector. The experimental objectives were:

1. To determine the relationship between the accuracy of the target detector and the standard deviation of the DOG filter used during contour tracing. This was important because it is reported in [Cumani, 1991] that there is an optimum value of standard deviation that provides maximum accuracy.

2. To compare the actual accuracy obtained with the developed target detector with that reported in [Cumani, 1991] and hence establish the effect of the differences between the two algorithms.

3. To ensure that the accuracy of the developed target detector was sufficiently high to meet the calibration requirements of the ASP stereo vision system under actual imaging conditions. i.e. in the presence of image noise and predicted levels of contrast between the target circles and the background.

10.3.5.2. Method

The first requirement was to generate a synthetic image of a circle with a centre at a known sub-pixel location and a size comparable to that obtained in images of the calibration grid. This was done by first using the HIPS `calcpix` filter to generate a 512×512 pixel image comprising a dark grey background (grey-level of 64) and a light grey circle (grey-level of 191) with a radius of 192 pixels. This image was then reduced by a factor of 4 using the HIPS `reduce` filter so that by moving the centre of

the original circle in integer pixel steps, the centre of the reduced circle could be controlled in steps of 0.25 pixels. For example, the following command line was used to generate a reduced circle from an original one centred at (256, 256):

```
fcalcpix -c 512 512 -s "if (((c - 256) * (c - 256) + (r - 256) \
      * (r - 256)) < (192 * 192)) opix = 191; else opix = 64" | \
      reduce -s 4 > circle.hips
```

This series of operations results in a 128 × 128 pixel image of a circle of radius 48 pixels and centre (63.625, 63.625). Note that the centre of the circle is not at the intuitively expected position of (64.0, 64.0) because image scaling is effectively centred at the location (-0.5, -0.5) and the image co-ordinates must therefore be translated appropriately before and after scaling. Consequently, the position of a point (x, y) in an image after scaling by a factor of S becomes:

$$(x,y) \xrightarrow{\text{Scaling by } S} \left(S\left(x+\tfrac{1}{2}\right)-\tfrac{1}{2}, S\left(y+\tfrac{1}{2}\right)-\tfrac{1}{2}\right) \qquad (10.12)$$

Two experiments were carried out using the synthesised circle image: the first to determine the optimum standard deviation for the DOG filter for different amounts of image noise and the second to determine the relationship between target detector accuracy for different amounts of image noise and contrast. In both experiments different amounts of sensor noise were simulated by adding Gaussian noise with standard deviations of ±1.0, ±1.5, ±2.0 and ±2.5 grey-levels to the basic circle image using the HIPS `gnoise` filter. A sequence of 100 different images was generated for each level of noise by using different random number seeds in a similar manner to that used in Chapter 7.

In the first experiment, the RMS error in the estimated position of the centre of the circle was obtained for each sequence of images when a range of standard deviations (sigmas) from 1.0 to 8.0 were used in the DOG filter. In the second experiment, different levels of contrast between the circle and background ranging from only 8 grey-levels up to the original 128 were obtained by linear scaling of the basic circle image before the noise was added. The RMS error in the estimated position of the centre of the circle was obtained for each level of noise and also for each level of contrast.

10.3.5.3. Results

The results obtained for the two experiments are shown in the graphs that are plotted in figure 10.7 and figure 10.8 respectively. Figure 10.7 shows the RMS error in the

detected centre of the circle plotted against the standard deviation of the DOG filter used during contour detection for a range of standard deviations of image noise and a contrast of 128 grey-levels between the circle and background. Figure 10.8 shows the RMS error in the detected centre plotted against the contrast between the circle and background for a range of standard deviations of image noise and a DOG standard deviation of 2.0 pixels.

10.3.5.4. Conclusions

It can be seen from figure 10.7 that the use of a DOG filter with a standard deviation of ±2.0 pixels results in the greatest accuracy for all levels of image noise. However, the graph also shows that that the RMS error of the detected position of the circle centre is actually quite insensitive to the standard deviation used. The accuracy achieved using a DOG filter with a sigma of 1.0 pixels is only approximately 10% worse than that obtained using one with a sigma of 2.0 pixels. Furthermore, a DOG filter with a sigma of 1.0 pixels requires a convolution mask that is only half of the size of that required for one with a sigma of 2.0. Therefore, unless extremely high accuracy is required, a DOG filter with a sigma of 1.0 pixels is probably preferable because it can be computed twice as quickly with minimal loss of accuracy.

Figure 10.8 shows that, as expected, the RMS error in the detected position of the circle increases both as the amount of noise increases (since this clearly degrades the signal-to-noise ratio) and also as the contrast between the circle and the background decreases (i.e. effectively reducing the amount of signal and hence also degrading the signal-to-noise ratio). For the levels of image noise and contrast obtainable in practice, an accuracy better than 0.01 pixels RMSE would be expected. However during calibration it may be desirable to use frame integration to increase the signal-to-noise ratio in images of the calibration grid and hence obtain a corresponding increase in the accuracy of the detected target locations.

The accuracy results obtained for the developed target detector compare quite favourably with those presented in [Cumani, 1991]. However it is not possible to perform a rigorous comparison of the accuracy of the two methods since details such as the size and contrast of the synthetically generated targets used to test the original algorithm are not reported in the paper.

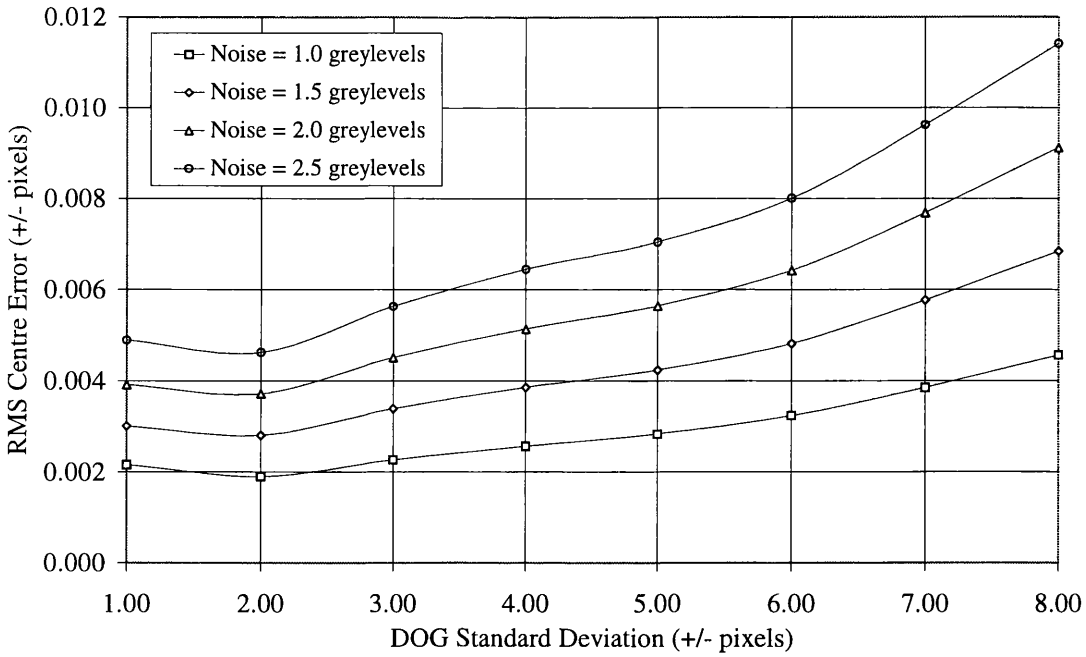


Figure 10.7. Graph of RMS error in the detected centre of the circle against the standard deviation of the DOG filter used during contour detection, for a range of standard deviations of image noise.

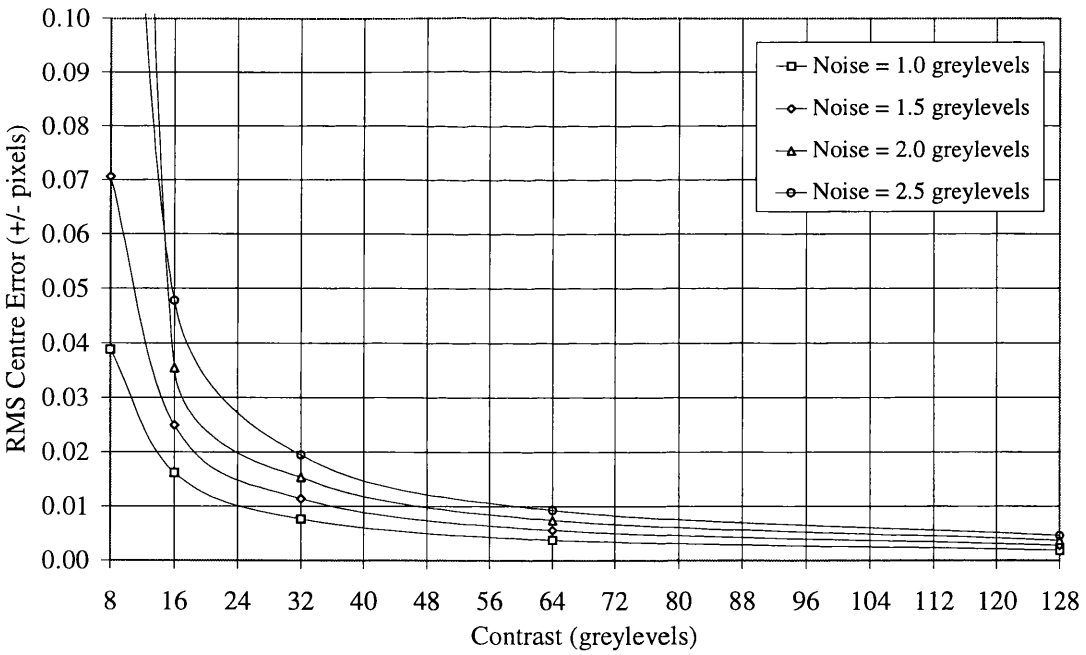


Figure 10.8. Graph of RMS error in the detected centre of the circle against the contrast between the circle and background.

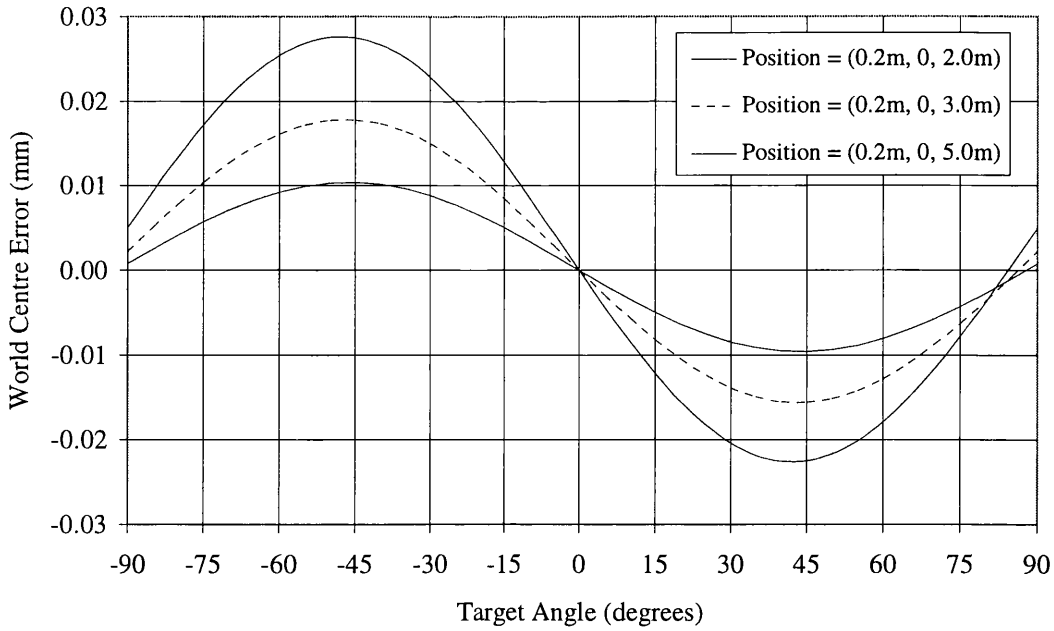


Figure 10.9. Error in the estimated world position of a circular target (of diameter 20mm) caused by perspective distortion against the relative angle of the target for various positions of the target with respect to the camera.

10.3.6. Other sources of error

The analysis of the previous section only considers the accuracy of the target detection algorithm itself when the image capture system is subject to noise. When the physical calibration grid is imaged, several additional sources of error will be introduced into the target detection process:

10.3.6.1. Perspective distortion

Unless the plane of a physical circular target is perpendicular to the imaging plane, it will be imaged as an ellipse and not as a circle as is assumed in the target detector. This, in itself, will not effect the accuracy of target localisation because the calculation of the centroid of a contour as an estimate of its centre is equally valid for an ellipse as it is for a circle. However, the distorting effect of perspective projection means that the centre of the imaged ellipse is not exactly that of the projection of the centre of the target circle. Fortunately, the results of simulation experiments performed by the author, such as those illustrated in figure 10.9, show that the error that is introduced by perspective distortion is extremely small for the target circle dimensions and positions used in the ASP stereo vision system.

10.3.6.2. Calibration grid illumination

The light source used to illuminate the calibration grid may lead to errors in the estimated position of the target circles if it is either non-uniform, or causes specular highlights. In [Cumani, 1991] the effects of a rather severe illumination gradient across a synthetic target image were investigated and found to be very limited.

However specular highlights have been observed to result in quite noticeable distortion in the detected target contours. Although the exact effect has not yet been measured, it seems clear that specular highlights may lead to significant errors in the calculation of the target centres. The specular highlights occur at the edges of the target because it is difficult to apply the paint right to the edge of the targets without some spilling over onto the body of the cylinder. Therefore, the reflective aluminium surface may show through at the edges of the targets, resulting in the observed specular highlights. A more effective method of obtaining the required contrast between the targets and the background than simple painting is required if specular highlights are to be avoided.

10.4. DLT-based static calibration

10.4.1. Introduction

Having detected the image locations of target circles and matched them with their corresponding world points, the final stage of static calibration is to determine the internal and external orientation parameters for the cameras. This then allows space intersection to be carried out to recover the world positions of matched stereo pairs of image points. The direct linear transform (DLT) [Abdel-Aziz, 1971] was the first method chosen for investigation because it is comparatively simple, it requires no initial estimates of the camera parameters and the requirement for different scale factors in the x and y directions of the CCD sensor is inherently accommodated in the method [Burner 1985].

10.4.2. Derivation

The derivation of the DLT starts with the basic definition of the collinearity equations given in Chapter 2. These are modified to incorporate an arbitrary affine transformation of the image co-ordinates that can incorporate the transformation caused by pixel sampling of the image. The collinearity equations therefore become (overleaf):

$$\alpha_1 x_i + \alpha_2 y_i + \alpha_3 = -c\frac{m_{11}(X_i - X_0) + m_{12}(Y_i - Y_0) + m_{13}(Z_i - Z_0)}{m_{31}(X_i - X_0) + m_{32}(Y_i - Y_0) + m_{33}(Z_i - Z_0)} \tag{10.13}$$

$$\alpha_4 x_i + \alpha_5 y_i + \alpha_6 = -c\frac{m_{21}(X_i - X_0) + m_{22}(Y_i - Y_0) + m_{23}(Z_i - Z_0)}{m_{31}(X_i - X_0) + m_{32}(Y_i - Y_0) + m_{33}(Z_i - Z_0)} \tag{10.14}$$

where: (X_i, Y_i, Z_i) are the world co-ordinates of a point $\mathbf{P}_i$; (x_i, y_i) are the observed (CCD frame) image co-ordinates of $\mathbf{P}_i$; c is the camera constant; m_{jk} are the elements of the orthogonal rotation matrix between the world and image co-ordinate systems; (X_0, Y_0, Z_0) are the object co-ordinates of the camera's perspective centre and $\alpha_1, \alpha_2, \ldots, \alpha_6$ are the coefficients of the affine transformation of the image co-ordinates. Abdel-Aziz and Karara [Abdel-Aziz, 1971] showed that equations (10.13) and (10.14) can be rewritten as the following pair of equations:

$$x_i = \frac{\beta_1 X_i + \beta_2 Y_i + \beta_3 Z_i + \beta_4}{\beta_9 X_i + \beta_{10} Y_i + \beta_{11} Z_i + 1} \tag{10.15}$$

$$y_i = \frac{\beta_5 X_i + \beta_6 Y_i + \beta_7 Z_i + \beta_8}{\beta_9 X_i + \beta_{10} Y_i + \beta_{11} Z_i + 1} \tag{10.16}$$

These DLT equations represent a linear relationship between the image and world co-ordinates in terms of the eleven DLT parameters, $\beta_1, \beta_2, \ldots, \beta_{11}$.

10.4.3. Implementation

10.4.3.1. Calibration

The objective of system calibration using the DLT is to determine the values of the eleven DLT parameters, $\beta_1, \beta_2, \ldots, \beta_{11}$, for a camera. If, as in the static calibration module, the world and image co-ordinates of a number of points are known for one camera, then equations (10.15) and (10.16) can be rearranged with the DLT parameters as unknowns to give:

$$x_i = \beta_1 X_i + \beta_2 Y_i + \beta_3 Z_i + \beta_4 - \beta_9 x_i X_i - \beta_{10} x_i Y_i - \beta_{11} x_i Z_i \tag{10.17}$$

$$y_i = \beta_5 X_i + \beta_6 Y_i + \beta_7 Z_i + \beta_8 - \beta_9 y_i X_i - \beta_{10} y_i Y_i - \beta_{11} y_i Z_i \tag{10.18}$$

Combining these equations for all $i = 1,2,\cdots,N$ points with known image and world co-ordinates gives a matrix equation of the following form:

$$\mathbf{A}\beta = \mathbf{b} \tag{10.19}$$

where (overleaf):

$$\mathbf{A} = \begin{bmatrix} X_1 & Y_1 & Z_1 & 1 & 0 & 0 & 0 & 0 & -x_1X_1 & -x_1Y_1 & -x_1Z_1 \\ 0 & 0 & 0 & 0 & X_1 & Y_1 & Z_1 & 1 & -y_1X_1 & -y_1Y_1 & -y_1Z_1 \\ X_2 & Y_2 & Z_2 & 1 & 0 & 0 & 0 & 0 & -x_2X_2 & -x_2Y_2 & -x_2Z_2 \\ 0 & 0 & 0 & 0 & X_2 & Y_2 & Z_2 & 1 & -y_2X_2 & -y_2Y_2 & -y_2Z_2 \\ \vdots & & & & & & & & & & \vdots \\ X_i & Y_i & Z_i & 1 & 0 & 0 & 0 & 0 & -x_iX_i & -x_iY_i & -x_iZ_i \\ 0 & 0 & 0 & 0 & X_i & Y_i & Z_i & 1 & -y_iX_i & -y_iY_i & -y_iZ_i \\ \vdots & & & & & & & & & & \vdots \\ X_N & Y_N & Z_N & 1 & 0 & 0 & 0 & 0 & -x_NX_N & -x_NY_N & -x_NZ_N \\ 0 & 0 & 0 & 0 & X_N & Y_N & Z_N & 1 & -y_NX_N & -y_NY_N & -y_NZ_N \end{bmatrix} \tag{10.20}$$

$$\beta = [\beta_1 \ \beta_2 \ \beta_3 \ \beta_4 \ \beta_5 \ \beta_6 \ \beta_7 \ \beta_8 \ \beta_9 \ \beta_{10} \ \beta_{11}]^T \tag{10.21}$$

$$\mathbf{b} = [x_1 \ y_1 \ x_2 \ y_2 \ \cdots \ x_i \ y_i \ \cdots \ x_N \ y_N]^T \tag{10.22}$$

An estimate of β, the vector of DLT parameters, can then be found from the least squares solution to equation (10.19):

$$\beta = (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{b} \tag{10.23}$$

Note that at least six non-coplanar points ($N \geq 6$) are required to ensure that the matrix $\mathbf{A}^T\mathbf{A}$ in equation (10.23) has full rank and can hence be inverted. A computer program to perform DLT calibration was implemented. This used equation (10.23) to derive the DLT parameters for the left or right camera, $\beta_{L1}, \beta_{L2}, \cdots, \beta_{L11}$ and $\beta_{R1}, \beta_{R2}, \cdots, \beta_{R11}$ respectively, from the detected locations of the target points in the left or right image and their known world locations.

10.4.3.2. Space intersection

If static calibration has been performed to find the DLT parameters for both cameras in a stereo vision system, then the DLT can be used to perform space intersection as long as the camera settings are not altered. After calibration, the DLT parameters are known and, after stereo matching, the left and right image points, (x_L, y_L) and (x_R, y_R), corresponding to the same world point are known. This forms the basis of the following derivation of space intersection using the DLT that was developed by the author. (However it should be noted that several authors have previously published similar derivations.) One set of equations (10.15) and (10.16) can be written for each camera and rearranged with the co-ordinates of the world point, (X,Y,Z), as unknowns (overleaf):

$$(\beta_{L9}x_L - \beta_{L1})X + (\beta_{L10}x_L - \beta_{L2})Y + (\beta_{L11}x_L - \beta_{L3})Z = (\beta_{L4} - x_L) \quad (10.24)$$

$$(\beta_{L9}y_L - \beta_{L5})X + (\beta_{L10}y_L - \beta_{L6})Y + (\beta_{L11}y_L - \beta_{L7})Z = (\beta_{L8} - y_L) \quad (10.25)$$

$$(\beta_{R9}x_R - \beta_{R1})X + (\beta_{R10}x_R - \beta_{R2})Y + (\beta_{R11}x_R - \beta_{R3})Z = (\beta_{R4} - x_R) \quad (10.26)$$

$$(\beta_{R9}y_R - \beta_{R5})X + (\beta_{R10}y_R - \beta_{R6})Y + (\beta_{R11}y_R - \beta_{R7})Z = (\beta_{R8} - y_R) \quad (10.27)$$

A solution for (X, Y, Z) can again be found from the least squares matrix equation:

$$\mathbf{x} = \left(\mathbf{C}^{\mathrm{T}}\mathbf{C}\right)^{-1}\mathbf{C}^{\mathrm{T}}\mathbf{d} \quad (10.28)$$

where:

$$\mathbf{x} = \begin{bmatrix} X & Y & Z \end{bmatrix}^{\mathrm{T}} \quad (10.29)$$

$$\mathbf{C} = \begin{bmatrix} (\beta_{L9}x_L - \beta_{L1}) & (\beta_{L10}x_L - \beta_{L2}) & (\beta_{L11}x_L - \beta_{L3}) \\ (\beta_{L9}y_L - \beta_{L5}) & (\beta_{L10}y_L - \beta_{L6}) & (\beta_{L11}y_L - \beta_{L7}) \\ (\beta_{R9}x_R - \beta_{R1}) & (\beta_{R10}x_R - \beta_{R2}) & (\beta_{R11}x_R - \beta_{R3}) \\ (\beta_{R9}y_R - \beta_{R5}) & (\beta_{R10}y_R - \beta_{R6}) & (\beta_{R11}y_R - \beta_{R7}) \end{bmatrix} \quad (10.30)$$

$$\mathbf{d} = \begin{bmatrix} (\beta_{L4} - x_L) & (\beta_{L8} - y_L) & (\beta_{R4} - x_R) & (\beta_{R8} - x_R) \end{bmatrix}^{\mathrm{T}} \quad (10.31)$$

A new HIPS filter to perform space intersection on a disparity map was written. This command takes the row and column position of each pixel in the disparity map as the left image co-ordinates, (x_L, y_L), of a point. The right image co-ordinates, (x_R, y_R), of the point are calculated by adding the vertical and horizontal disparity estimates for the pixel to its row and column co-ordinates. The 3D location of the point is then obtained by inserting the two sets of image co-ordinates and the left and right camera DLT parameters into equation (10.28).

It should be noted that a solution to equation (10.28) requires that the matrix $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ has full rank and can hence be inverted. Although this is not guaranteed (e.g. consider the extreme case $\beta_{Li} = \beta_{Ri}, i = 1,2,3,9,10,11$, $x_L = x_R$, $y_L = y_R$), the HIPS filter that was written to perform DLT-based space intersection returns an error code for any points for which the determinant of $\mathbf{C}^{\mathrm{T}}\mathbf{C}$ is found to be zero. However, it is also noteworthy that this situation has never arisen in practice using the ASP sensor system.

10.4.4. Accuracy analysis

10.4.4.1. Objective

The objective of this experiment was to analyse the accuracy, robustness and consistency of the complete static calibration process, from the capture of images of the calibration grid to the detection of the target points and subsequent calibration using the DLT.

10.4.4.2. Method

The ASP stereo vision system was set-up with a baseline of approximately 300mm and verged approximately to the world point (0, 0, 2.0m). The cameras were mounted in "portrait orientation", i.e. with the cameras' scan-lines running vertically from top to bottom. This orientation of the cameras was found to provide images with an aspect ratio better suited to the capture of human faces (and the calibration grid) than the more conventional "landscape orientation".

The calibration grid was placed in the field of view of the cameras so that it was centred roughly at the vergence point. Fifty stereo images of the calibration grid were then captured consecutively with approximately 30 seconds between captures. The ASP sensor head and the calibration grid were placed on the same bench while the images were being captured in order to minimise the inevitable relative movement between them caused by vibration.

The target detection algorithm that was described in section 10.3 was used to locate the centres of the target circles in both the left and right images of all 50 of the captured stereo images. The results of target detection are given in sub-section 10.4.4.3. Each of the 50 sets of detected target image co-ordinates were combined with the nominal world positions of the corresponding target points in equation (10.23) to derive 50 different estimates of the DLT parameters for the left and right cameras. Two measurements of the accuracy of the calibration were then calculated: the *image error*, which is described in sub-section 10.4.4.4, and the *world error*, which is described in sub-section 10.4.4.5.

10.4.4.3. Target Detection

An example of one of the captured left images and the image locations of the target centres that were subsequently detected are shown in figure 10.6. The same 28 targets that were detected in this image were also detected in all of the other 49 left images. However, one additional target that was not detected in the left images (the top right

hand one) was detected in the right images. Although this additional right image point was included during the derivation of the right image DLT parameters, for simplicity it was ignored during the calculations of image error and world error that are described in sub-sections 10.4.4.4 and 10.4.4.5 respectively.

The spread of the detected target image points was measured so that the consistency and accuracy of the target detector could be estimated. This was done by calculating the deviation from the mean position of each point in all 50 of the captured left or right images. Figure 10.10 shows a graph of the deviation from the mean of the detected row and column position of one particular point. The total RMS deviation from the mean of the row and column positions of all the detected points in all of the captured left and right images is given in table 10.4.

Image	RMS deviation of location (pixels)		Absolute maximum deviation of location (pixels)	
	Row	Column	Row	Column
Left	±0.080	±0.010	0.241	0.048
Right	±0.083	±0.010	0.250	0.034

Table 10.4. Total RMS and absolute maximum deviation from the mean of the row and column positions of all points.

10.4.4.4. Image Error

Each set of detected target image co-ordinates was combined with the nominal world positions of the targets in equation (10.23) to derive a separate estimate of the DLT parameters for each captured image. The world co-ordinates of the target points were then projected onto the left and right image planes by inserting the obtained DLT parameters into equations (10.15) and (10.16). This resulted in a separate set of *predicted image points* for each of the 50 sets of DLT parameters obtained for the left and right cameras. The image error for each point is the vector difference between the row and column co-ordinates of the model predicted image point and those of the corresponding originally observed image point.

Figure 10.11 shows the mean image error vector for each point plotted at 100 times magnitude from the mean observed location of the target point. The total RMS error in the row and column positions of the detected targets, calculated over all 50 left or right images and all 28 targets, is given in table 10.5. The average standard deviation of the error for each point is also shown in the table.

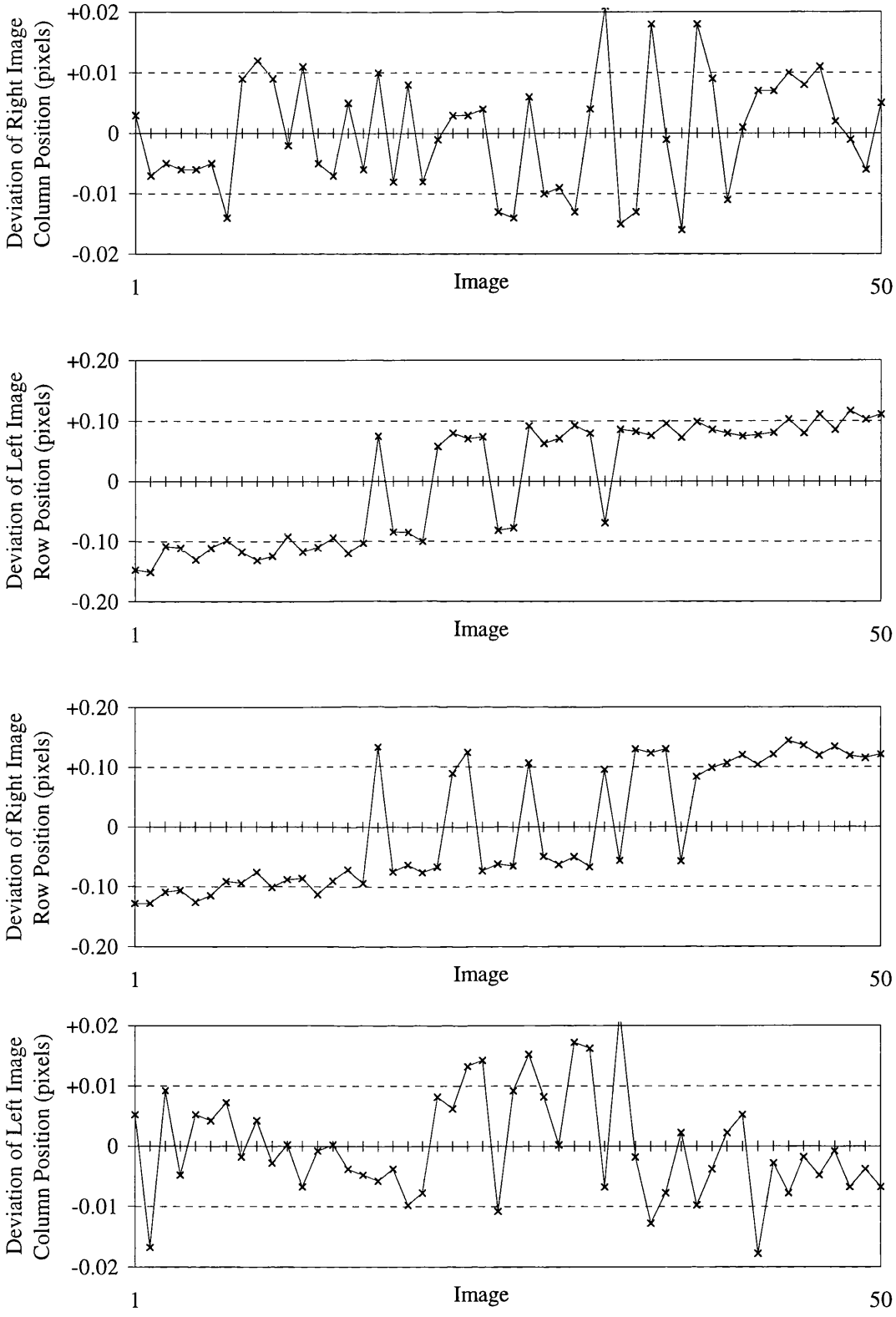
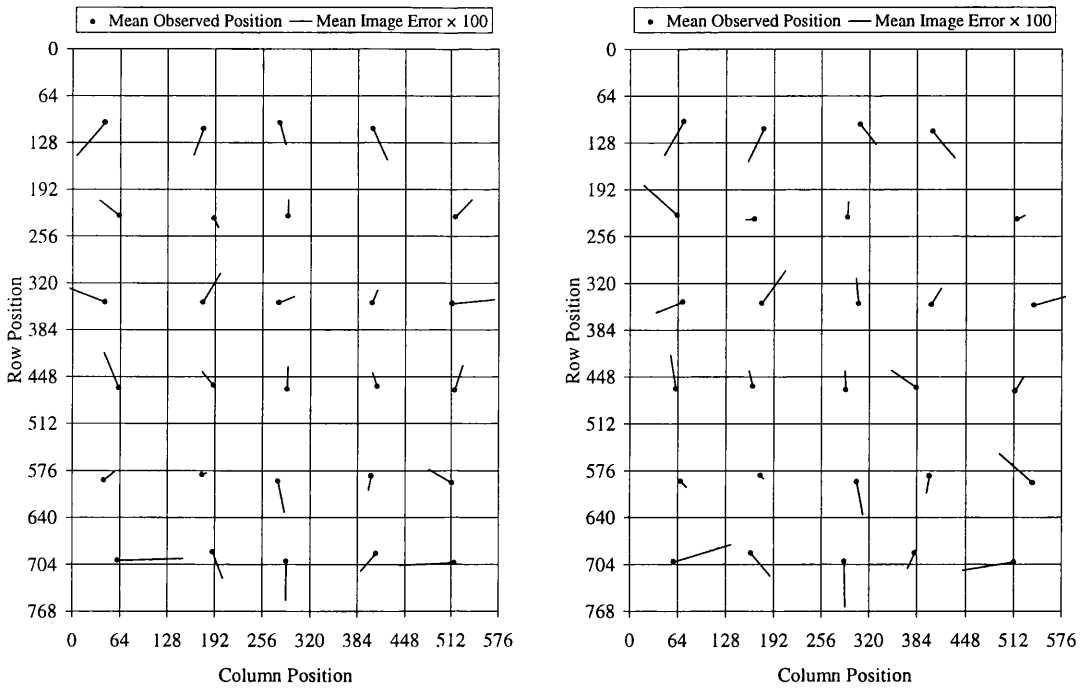


Figure 10.10. Deviation from the mean of the detected column and row positions of a target point in the right and left images. (Note that different scales were used to plot the row and column positions).



Left image. *Right image.*

Figure 10.11. The mean image error vectors (plotted at 100× magnitude).

Image	RMS image error (pixels)		Standard deviation (pixels)	
	Row	Column	Row	Column
Left	±0.294	±0.316	±0.080	±0.080
Right	±0.323	±0.312	±0.089	±0.074

Table 10.5. Total RMS image error and the average standard deviation of the image error for each point.

10.4.4.5. World Error

The co-ordinates of the *predicted world points* were obtained by space intersection of the stereo pairs of observed image points using equation (10.28) and the appropriate sets of DLT parameters. The difference between the co-ordinates of the predicted world points and the original world co-ordinates of the target points is the *world error*. This indicates how closely the derived DLT model fits the actual physical imaging system in terms of the world co-ordinates. Table 10.6 shows the total RMS world error for all 28 target points that were detected in all 50 stereo pairs of images.

	X	Y	Z
RMS world error (mm)	±0.106	±0.101	±0.227
Standard deviation (mm)	±0.003	±0.010	±0.028

Table 10.6. Total RMS world error and the average standard deviation of the world error for each target point.

10.4.4.6. Interpretation of results

It can be seen from the results given in table 10.4 and figure 10.10 that the image column locations provided by the target detector have been highly consistent, with a standard deviation of only ±0.01 pixels. However the results also show that the detected row position exhibits significantly higher variation, with a standard deviation of ±0.080 pixels for the left image and ±0.083 for the right image.

The probable cause for the difference in variation between the detected row and column positions of the target points is timing jitter between the cameras and the frame-store. This introduces a timing error in each scan-line of the image signal that effectively results in the addition of noise to the column position of pixels when the image is reconstructed in the frame-store. In this experiment, the cameras were arranged in "portrait orientation" with the images being rotated by 90° after capture. Hence timing jitter will effect the row position and not the column position of pixels in the final rotated images.

Nevertheless, table 10.5 shows that the RMS image error is approximately ±0.3 pixels for both the row and column positions while the average standard deviation of the image error for each point is much lower. Therefore, the higher variation in the row position of the detected target does not appear to affect the accuracy of the DLT calibration significantly. The low variance of the image error for each point would appear to imply that the errors in the DLT calibration are mainly systematic.

The broad geometric distribution of the image errors shown in figure 10.11 and the fact that the errors are very similar for both cameras suggests that they may be due to errors in the world co-ordinates of the target points. This is quite possible since only the nominal positions of the target circles taken from the drawings was used in the DLT calibration.

Despite the presence of the systematic errors described above, a high level of accuracy has still been achieved using the DLT. It can be seen from table 10.6 that the X- and Y-

components of the RMS world error are approximately ±0.10mm, while the *Z*-component is approximately ±0.23mm.

10.5. Conclusions

The DLT-based approach to static calibration that has been investigated in this chapter has been largely successful in meeting the original objectives. The developed static calibration module has been shown to be capable of both calibrating the ASP stereo vision system and subsequently performing space intersection to recover the 3D surface from the disparity map derived from a captured stereo pair of images.

Although it was not possible to measure the accuracy of recovered 3D surfaces, the obtained results give encouragement that the required accuracy of ±1.0mm RMSE can be obtained. The analysis described in sub-section 10.3.4 shows that the target detection component of static calibration can provide a level of accuracy that is well in excess of the requirements. In sub-section 10.4.4, it was also shown that a DLT model can be derived that fits the physical imaging system with an RMS error of ±0.10mm in *X* and *Y* and ±0.23mm in *Z*.

As well as going a considerable way to meeting the primary accuracy requirement of static calibration, the developed system also meets the secondary requirements of ease of use and quick operation. The only part of the process that is not performed completely automatically by computer is matching the image and world target points. However this task has been made very straightforward through the implementation of a suitable graphical user interface. The target detection process takes approximately 30 seconds on a Sun SPARCstation 20 for a full CCIR image, while the subsequent DLT calibration is performed in about one second. These times mean that the speed of the DLT-based static calibration process easily meets the requirements for interactive use.

Chapter 11

System integration and final results

11.1. Introduction

As stated at the beginning of chapter 10, a prototype videometrics system called CopyCAD was to be developed to demonstrate the basic components of the ASP stereo vision system (acting only in a static mode of operation) applied to the problem of maxillo-facial measurement. It was explained in chapters 2 and 3 (see also figure 3.1) that the main components of such a videometrics system are: stereo image acquisition, stereo correspondence, system calibration (including space intersection) and 3D visualisation.

The work described in the preceding chapters has been aimed at the separate development of each of the above components of a videometrics system in isolation. In the work described in chapters 4 to 7, the ASP sensor system was developed specifically to provide a highly flexible stereo image acquisition facility. A version of the MSSM stereo matcher that was proven to be highly effective at solving the stereo correspondence problem resulted from the re-development and performance characterisation of the algorithm that are described in chapters 8 and 9 respectively. Finally, chapter 10 described the development and implementation of a DLT based approach developed that was shown to successfully address the system calibration problem.

Therefore, the one remaining problem, following the development of separate components for image acquisition, stereo correspondence and system calibration, was their integration into a single complete videometrics system. Inherent in this problem is the essential requirement that a user interface be provided to control the integrated system. Section 11.2 of this chapter describes the software and graphical user interface that were developed to address this problem, integrate the three main components, and hence produce the final prototype of the CopyCAD system. Results of 3D surface

recovery obtained using the completed CopyCAD system are presented in section 11.3.

11.2. Software integration and user interface design

11.2.1. Overview

Figure 11.1 shows a schematic of the structure of the main software components in the CopyCAD videometrics system. This comprises five main modules: sensor head control including image capture, system calibration, the stereo matcher, space intersection and 3D computer graphics. An X-Windows user interface, integrating all of these modules, was developed for the host workstation using Tcl and Tk [Ousterhaut, 1994]. Details of each of the five modules is given below.

11.2.2. Sensor head control

The sensor head control module was designed to allow the user full control over all aspects of the ASP sensor system. It was written mostly in standard Tcl with the addition of some custom developed Tcl commands, written in C, to interface with the drivers that had been developed for the ASP sensor head. The main functions of the sensor head control module are:

- *Projector control*: provides control of the projector focus and the slide advance mechanism. This allows the current slide to be selected either manually or automatically from: a red, green or blue filter for colour capture, a texture slide to capture stereo images for subsequent stereo matching or a neutral density filter for monochrome capture. The texture slide is maintained sharply in focus, while the others are de-focused to avoid projecting scratches or dust from the slide onto the subject.

- *Viewpoint control*: provides control of the azimuth and elevation position of the cameras and projection mirror. Each set of actuators can be moved independently or in tandem. An additional vergence and divergence control is also provided for the camera azimuth actuators.

- *Image preview*: allows the live images from the cameras to be displayed or frozen on the RGB monitor. The image from a single camera can be displayed in monochrome or the images from both cameras can be displayed as a red-green stereo anaglyph.

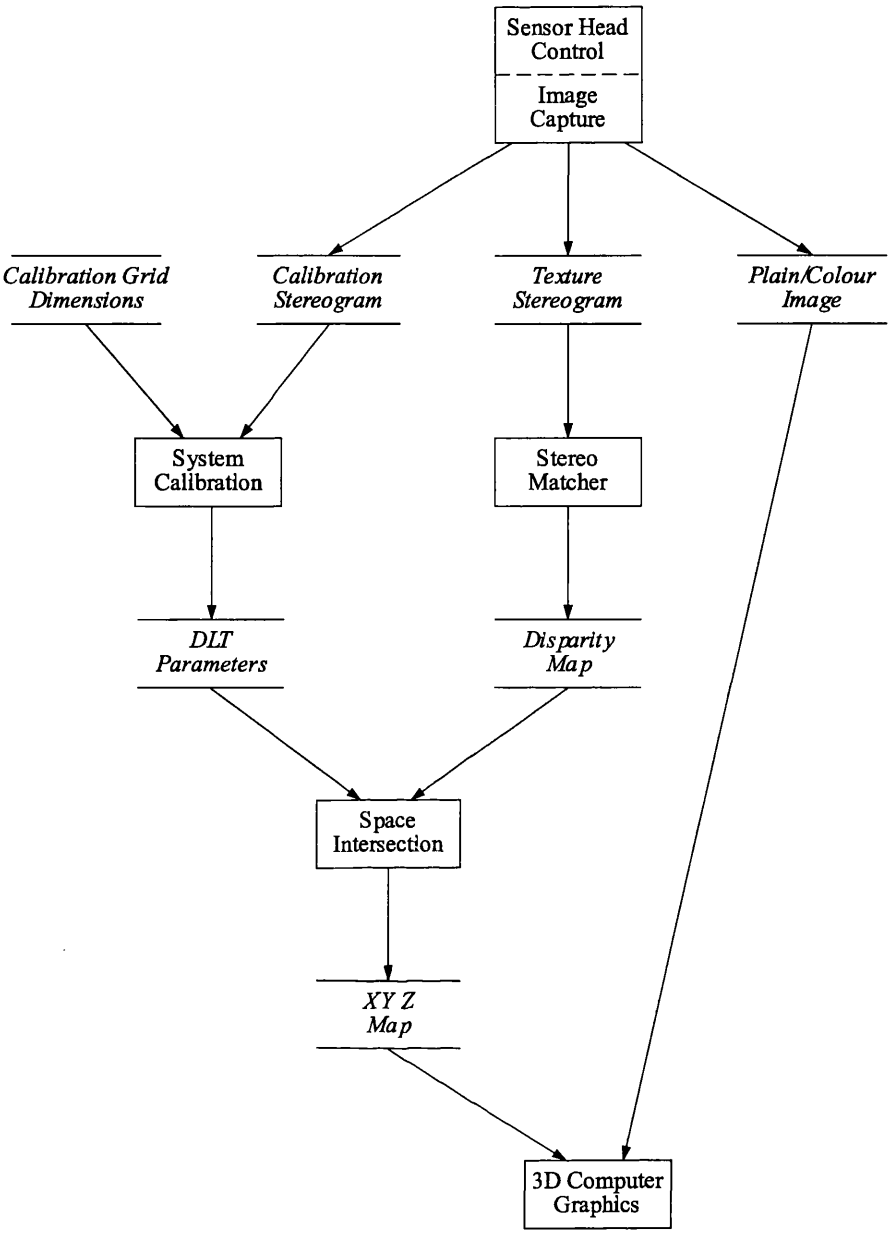


Figure 11.1. Schematic of the CopyCAD system.

- *Image capture*: allows image sequences to be captured automatically using any combination of camera and slide; e.g. both left and right images with the texture slide, the left image with the neutral density slide and the right image with the red, green and blue filter slides. Captured images are stored in memory and displayed as thumb-nail icons in the user interface.

- *Image selection*: allows previously captured images to be selected so that they can be passed onto one of the subsequent processing modules. Selected images can

also be displayed, deleted or saved to a file so that they can be subsequently loaded back into the system at a later date.

Figure 11.2 on page 226 shows a screen dump of CopyCAD's interface to the image acquisition module.

11.2.3. System calibration

The system calibration module controls all of the aspects of static system calibration that were described in chapter 10. Calibration requires that a stereo pair of images of the calibration grid be captured as described above. The user selects either the left or the right image and the selected image is then processed to determine the locations of the target circles that are visible within it. The interface to the system calibration module is shown in figure 11.3 on page 226. The main image panel in the figure is shown displaying the left image from a stereogram of the calibration grid. The contours of detected targets along with crosses at their estimated centres are overlaid in colour on top of the displayed image.

The array of coloured circles in the top left hand corner of the calibration module interface represents the world positions of the target points on the calibration grid. The co-ordinates of these points are read in from an external file. The user associates the world and image locations of target points by first "clicking" on a world point and then "clicking" on the corresponding contour in the main image display.

System calibration is carried out after the image locations of the targets have been detected and associated with their corresponding world points. This is done automatically using a new Tcl command, called `dlt`, that was developed to perform DLT calibration. The `dlt` command takes as input a Tcl list of world point co-ordinates and a second Tcl list containing the associated image point co-ordinates. The output of the command is a list of the eleven DLT parameters for the selected camera. The list of calculated DLT parameters is written out to a temporary file so that the parameters can be used by the space intersection module later in the session.

11.2.4. Stereo matching

The stereo matching module performs stereo matching using Version III of the MSSM stereo matcher as describe in chapter 8. The stereo matching module first requires that a stereo pair of images be captured, usually while the subject is bathed in textured light, and selected. The selected stereo images are then piped into a HIPS pipeline

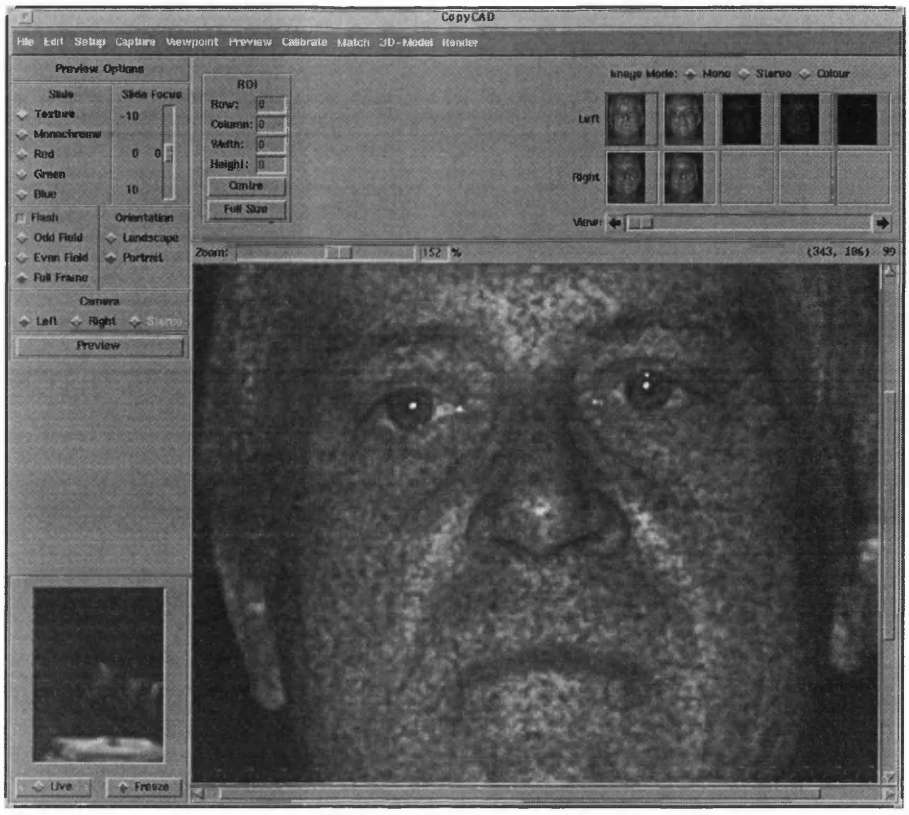


Figure 11.2. CopyCAD interface to the image acquisition module.

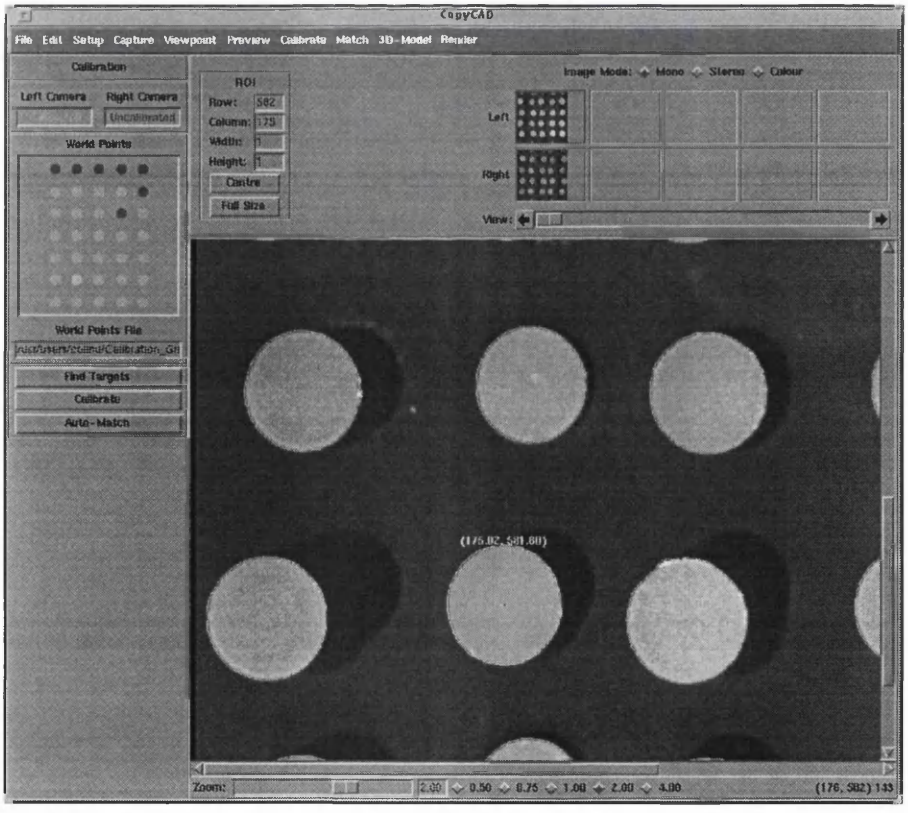


Figure 11.3. CopyCAD interface to the system calibration module.

process where stereo matching is performed. The output disparity and confidence maps are saved, in HIPS floating point format, to a user specified file. They are also scaled to HIPS byte format images that are then loaded back into memory and displayed to give the user a visual indication of the quality of the match obtained.

11.2.5. Space intersection

Space intersection is performed following stereo matching and after system calibration has been carried out for both cameras. The disparity map file and DLT parameter files are read into the specially developed HIPS filter, `dlt_dtor`, that was described in chapter 10. This filter performs space intersection, deriving the 3D world co-ordinates (*X, Y, Z*) for every point in the input disparity map. The output "XYZ" map is saved, in HIPS floating point format, to a user specified file. The output XYZ map is also displayed in the CopyCAD interface in a similar way to the disparity and confidence maps.

11.2.6. 3D computer graphics

A 3D computer graphics package was required to manipulate and display the 3D surface data contained in the XYZ map that is output by the space intersection module. The package initially used was one that was originally developed by Yau and Duffy at Heriot-Watt University [Yau, 1988; Yau 1989] and later re-implemented by Skelly at the Turing Institute [Skelly, 1992]. This package, in common with most 3D computer graphics system, first produces a polygon model from the input XYZ map. The polygon model is then used to generate an animation sequence of the recovered 3D surface. This can be displayed either as a wire-frame or smooth shaded (Gouraud or Phong) surface. If an additional plain image were available, such as a non-texture projected image captured at the same time as the texture stereogram, this could be used as a texture map to produce a photo-realistic render of the surface.

However the Yau/Skelly 3D computer graphics system suffers from the disadvantages that it was non-interactive, rather slow and quite inflexible. Therefore, when a Silicon Graphics computer became available, the Yau/Skelly system was replaced by Silicon Graphic's real-time, fully interactive, IRIS Inventor 3D computer graphics package. In order to interface CopyCAD to IRIS Inventor, another computer program was obtained to convert the HIPS format XYZ maps produced by CopyCAD to IRIS Inventor's polygon model format. The polygon model file is then read into IRIS Inventor where it can be manipulated and displayed interactively using a number of different surface finishes, including: wire-frame, smooth shaded and colour or

monochrome texture rendered. Examples of some rendered 3D surface models that were thus obtained are given in the following section.

11.3. 3D surface recovery

11.3.1. Objective

The objective of this experiment was to evaluate the ability of the CopyCAD system to capture 3D computer graphics models of the human face for maxillo-facial measurement applications. The primary factor to be determined was whether the system could be used, in practice, to recover 3D surface models of live human subjects' faces. It was also intended that other secondary factors, such as the overall ease of use of the system, would also be determined during the experiment.

11.3.2. Method

The CopyCAD system was used to capture a stereo pair of images of the calibration grid and both cameras were calibrated as described above. With the system then calibrated, CopyCAD was used to capture two stereo pairs of images of various subjects' faces: one stereo pair with the subject illuminated using textured light and the other with the subject illuminated using normal light. The texture illuminated stereo pair was then matched and a 3D surface model was generating by space intersection of the resulting disparity map using the previously obtained system calibration data. The IRIS Inventor package, running on a Silicon Graphics workstation, was used to display, manipulate and inspect the quality of the recovered 3D surface.

11.3.3. Results

Figure 11.4 shows an example of the central 512×320 pixel region of the stereo pairs of images captured for one particular subject (the author). Snapshots of the 3D surface model obtained for this region using the method described above are shown in figure 11.5. The figure shows the obtained surface displayed as a wire-frame model, a Gouraud shaded surface and a surface texture rendered with the left image from the normally illuminated stereo pair of images.

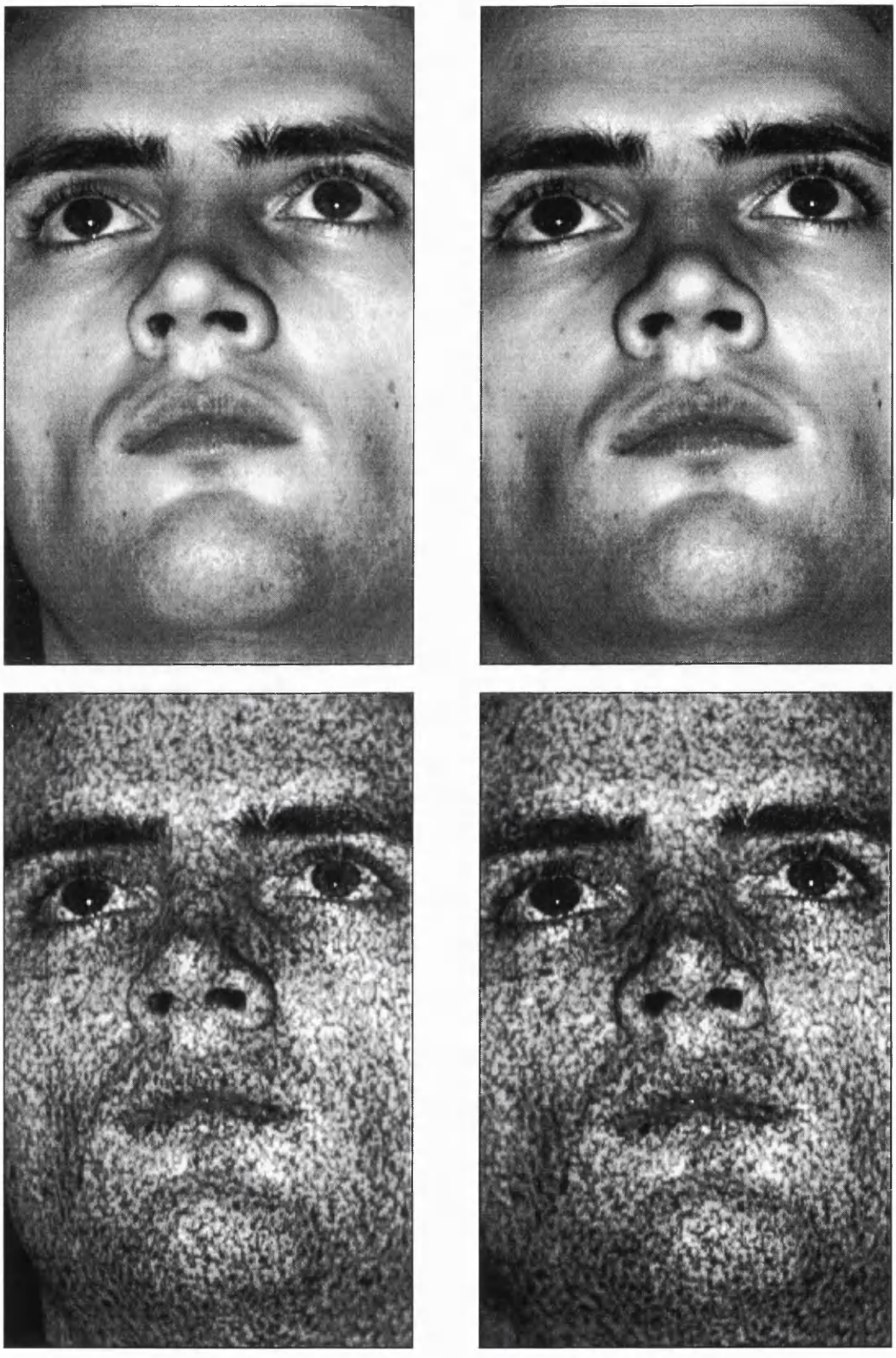


Figure 11.4. Input images. Top: stereo pair captured under normal illumination. Bottom: stereo pair captured under textured illumination.

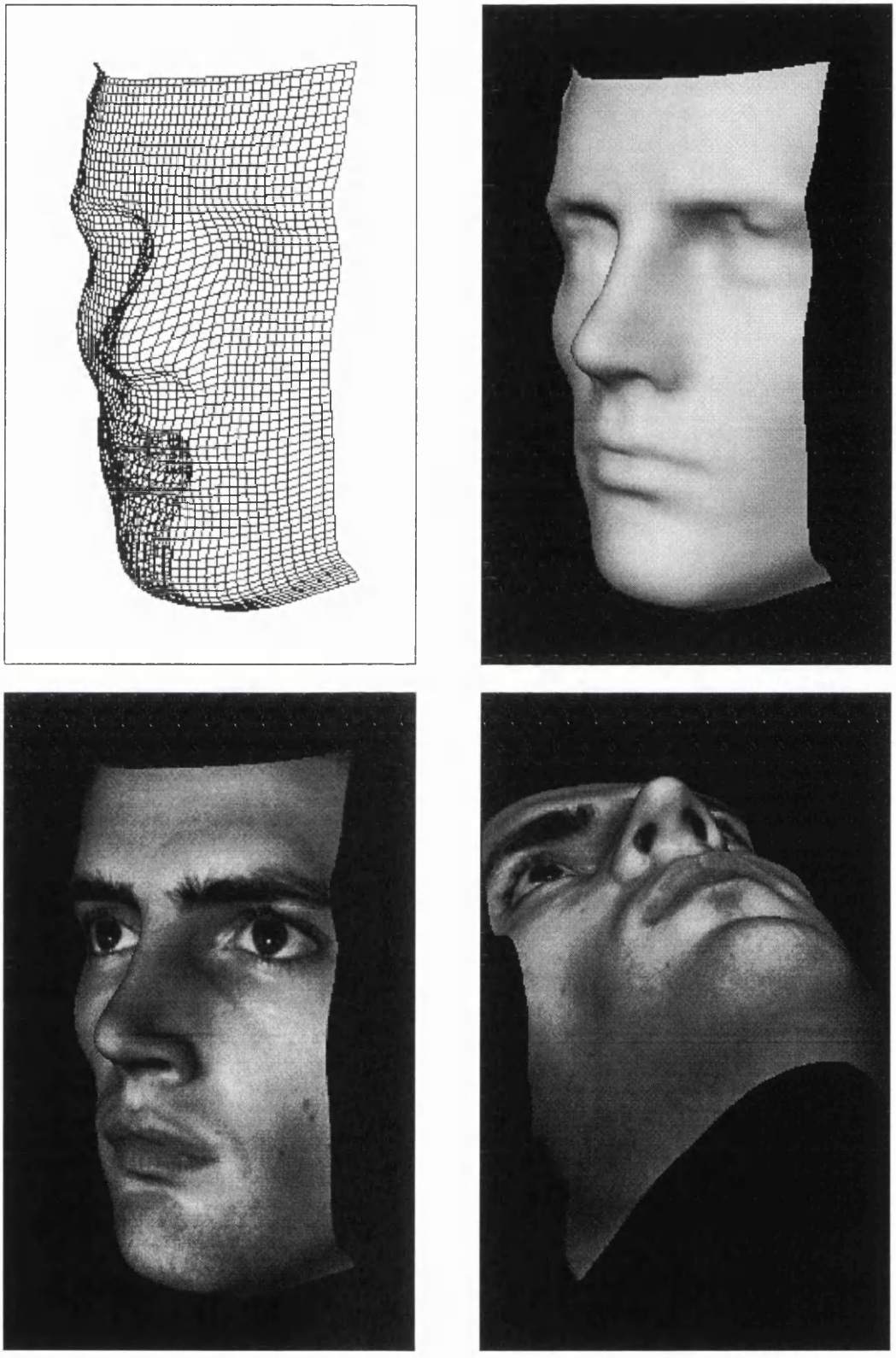


Figure 11.5. Recovered surface model. Top: wire-frame and Gouraud shaded. Bottom: texture rendered.

11.4. Conclusions

The results shown in figure 11.5 clearly demonstrate that the CopyCAD system is capable of recovering realistic 3D surface models of live human subjects' faces. The quality of these initial results appears to be entirely satisfactory for applications in the entertainment industry, such as the building of 3D face models for film, television or computer games. However, a full clinical trial would be required before the applicability of the system for use in more stringent medical applications, such as maxillo facial measurement, could be verified.

The approach adopted for the integration of the separate videometrics components together with a graphical user interface has proven to be very successful. The developed graphical user interface was found to allow 3D surface models to be captured in a simple and convenient manner. The fact that the system can be operated easily without any specialist photogrammetric training provides further evidence of its suitability for wider exploitation.

It would have been desirable to obtain a quantitative figure for the measurement accuracy of the system. However, this would have required that the CopyCAD system be used to measure a surface whose dimensions were already known to a high degree of accuracy. (In fact it is commonly accepted that the accuracy of the known values should be at least an order of magnitude greater than that of the measurement device under test.) It was believed that the only practical method for providing the required alternative measurements of a surface was to obtain an specially constructed test surface. Unfortunately sufficient resources were not available to provide such a test surface. The topic of quantifying the surface measurement accuracy of the complete CopyCAD system is discussed in greater depth in the Further Work section of chapter 12.

Chapter 12

Conclusions and further work

12.1. Review of thesis goals

At the outset of the work described in thesis, the overall goal was to develop the Active Stereo Probe (ASP), a dynamic stereo computer vision system capable of fully automatic, on-line, quantitative 3D surface reconstruction. At that time, the state-of-the-art in computer vision research was lacking significantly in three major areas essential to the successful achievement of this highly ambitious goal. One of the objectives of the research described in this thesis was therefore to advance the state-of-the-art in these three areas to such an extent that the development of the ASP system would be possible. The three areas of computer vision research that required advancement were:

1. Sensor system design and construction.

2. Automatic computer stereo matching.

3. Vision system calibration.

However, in addition to the basic research required in the above three areas, the problem of integrating the solutions developed in each area into a practically useful system was, significantly and quite uniquely, given a high priority. Indeed, one of the most novel aspects of this work was the great emphasis that was placed on the specification and development of a fully integrated system from the very outset. It is apparent from the previous chapters of this thesis that this approach has contributed greatly to the ultimate success of the work.

The aim of the remainder of this chapter is firstly to review the contributions made by this work to the advancement of the three areas of computer vision research listed above and to the wider problem of developing an integrated videometrics system.

Subsequent work carried out by the author and colleagues that is based directly on the results presented in this thesis, and that furthers the thesis' broad goals, will then be described. Possible directions for the future advancement of this work will also be discussed.

12.2. Achievements

12.2.1. Sensor system design

Prior to the work described in thesis, computer vision researchers had not, on the whole, given a high priority to the task of designing and constructing their sensor systems to perform a specific vision task. This was particularly true in the field of dynamic computer vision, where the prevailing attitude to sensor system design was one of: "let's just build something and then see what we can do with it". "Not very much!" appears to have been the predominant result of this approach. Very few demonstrations of dynamic vision systems performing even the simplest of useful tasks has been previously presented. This is despite the fact that dynamic vision has been theoretically proven to provide many advantages over static vision.

The main reason for the dearth of previous examples of the practical use of dynamic sensor "heads" appears to have been that the final objective of their development has been simply the fabrication of the head itself. Previously reported heads have not been designed specifically to achieve a given level of performance or to be applied to a particular vision task. As a result, the few dynamic vision systems that have actually been applied to visual tasks have been found to be too kinematically unstable or inaccurate to provide useful results, particularly for quantitative vision tasks.

In order to address the above problem, a novel alternative to previously reported approaches to dynamic computer vision system development was proposed and investigated in this work. The main premise of this new approach was to design the system from the very outset specifically to perform a given task, namely 3D surface recovery from stereo. This allowed the performance requirements of the system to be identified and specified in terms of the parameters of a stereo vision system, e.g. the desired 3D surface reconstruction accuracy, viewing volume, working volume and disparity range.

Having specified the performance requirements, it was then planned that the system would be designed to achieve this target level of performance. However, because this approach had not been previously attempted, neither the data required nor even a method to derive a design specification for the sensor head from the performance

requirements was available. Indeed it soon became apparent that it would not be a simple task to obtain this data due primarily to the complexity of the asymmetrically convergent stereo imaging geometry that was to be employed.

It was concluded that the only practical method for addressing this problem was to build a computer simulation of the vision system and the stereo imaging process. This was achieved by developing a mathematical model of the stereo imaging system which was then implemented within a software simulation. This model was utilised successfully to determine both a specification for the basic configuration of the sensor head (i.e. the choice of cameras, the stereo baseline, the mean object distance and the lens principal distance) and a more detailed specification for the accuracy and tolerances required of the head's components. Having derived this design specification, it was a relatively simple task to select appropriate components and fabricate the sensor system with a high degree of confidence that the completed system would meet the original performance specification. It is important to note that, in addition to providing the results required for the specific instance of the ASP sensor system, the work has also produced a general methodology for the design and specification of arbitrary videometric capture systems.

12.2.2. Computer stereo matching

12.2.2.1. The MSSM stereo matcher

Computer stereo matching was the second area that required further research in order to achieve the thesis goal of the development of a stereo computer vision system capable of fully automatic, on-line, quantitative 3D surface reconstruction. At the beginning of the work, no computer stereo matcher was available that fulfilled all of the specified requirements of the stereo matcher for the ASP system. However the MSSM stereo matcher was chosen as a basis for the development of a new stereo matcher specifically for the ASP system. The MSSM stereo matcher was chosen because its basic design already met most of the more stringent requirements of the ASP system, namely it was completely automatic, i.e. it did not require input of manual seed points, its output was a dense set of disparity estimates, one for each input pixel, and it was capable of sub-pixel disparity estimation. This combination of features, which would be essential if the requirements of the ASP system were to be met, was believed to be unique to the MSSM stereo matcher.

However, there was one serious impediment to the use of the MSSM stereo matcher in the ASP system: its generally slow speed of execution and, in particular, the fact that the time taken to match a stereo pair of images rose steeply with the expected

range of disparity in the images. This was of particular concern because it effectively reduced the dynamic range of disparity that could be handled by the matcher and hence resulted in reduced disparity resolution. A secondary reason for this being undesirable was that it implies that the user must estimate the maximum amount of disparity expected in the input imagery in order to obtain the best performance from the algorithm. Therefore, although previous experiments had shown that the MSSM stereo matcher could perform very effectively, particularly when used in conjunction with textured scene illumination, its slow speed of execution had consigned it to use only as an experimental tool.

12.2.2.2. Performance improvement

The primary objective of the research carried out by the author into computer stereo vision was therefore to investigate methods of addressing the above shortcomings in the performance of the basic MSSM stereo matcher. This research resulted in the development of Version II of the MSSM stereo matcher that successfully increased the algorithm's speed of execution by up to 100 times. This was done primarily by re-designing the algorithm to use highly computationally efficient scale-space pyramid image structures. Moreover, as well as increasing the MSSM stereo matcher's basic speed of execution, this pyramid image structure also greatly reduced the previous limitation on the maximum range of disparity acceptable in input stereo images. These improvements both made the use of the MSSM stereo matcher in real applications a practical proposition, and also laid the foundations for yet further improvements.

Building upon the foundations laid by the author in the development of Version II of the MSSM stereo matcher, Arthur van Hoff of the Turing Institute developed Version III of the algorithm. Van Hoff improved the implementation of the pyramid image structure that was used in Version II and made several other refinements to achieve a further increase in execution speed by up to 50 times.

Prior to this work, the restricted dynamic range and extremely long processing times made the original MSSM stereo matcher virtually unusable in a practical context. The two stages of development successfully increased the intrinsic speed of execution of the original MSSM stereo matcher by upwards of 1000 times and vastly improved its dynamic range. These improvements have resulted in a stereo matcher that can now be employed in real world applications. More specifically, Version III of the MSSM stereo matcher meets the requirement of the ASP system to match full CCIR sized images in just over one minute (using a Sun SPARCstation 10).

12.2.2.3. Performance characterisation

Before the MSSM stereo matcher could be built into an application system such as the Active Stereo Probe, it was necessary to obtain a characterisation of its performance to ensure that it would be adequate to meet the system's requirements. This was particularly important following the development of Version III of the algorithm since only very limited practical experience had been obtained in its use. Rather surprisingly, almost no details of previous attempts to characterise stereo matching algorithms have been reported in the computer vision literature. It was therefore necessary to devise and implement from first principles a series of stereo matcher performance characterisation experiments. Although these procedures were developed primarily to address the immediate requirements of the MSSM stereo matcher, they were also designed so that they could be used to characterise and compare the performance of any correlation-based stereo matchers.

A profile of the match accuracy of the MSSM stereo matcher against image noise for synthetically generated Gaussian noise stereograms was obtained from one of the characterisation experiments. This indicated that a match error of between ±0.1 and ±0.2 pixels RMSE might be expected for images with noises levels similar to those obtained from current high quality CCD cameras. Additional characterisation experiments demonstrated that the algorithm's performance is insensitive to factors such as integer and sub-pixel disparities. A final experiment characterised the degradation in the accuracy of the MSSM stereo matcher with positive and negative disparity gradients. It was concluded from this series of characterisation experiments that the MSSM stereo matcher was free from major faults and that its performance would fulfil the requirements of the ASP system.

12.2.3. Vision system calibration

The third major area of computer vision research explored in this thesis was vision system calibration. The dynamic nature of the ASP sensor system demanded that any calibration scheme used should also be dynamic, i.e. be capable of maintaining the calibration of the cameras when they move. None of the previously reported approaches to dynamic calibration have been shown to be capable of providing the level of accuracy required by the ASP system. Therefore a new scheme was proposed that combined elements of photogrammetry with the accurate sensor position feedback and high stability of the ASP sensor head. It was believed that this new scheme would be capable of providing both the speed of computation and accuracy required by the ASP system.

Previously, photogrammetric techniques have not been widely used in the computer vision field. Photogrammetric solutions have also traditionally been inflexible, difficult to use and highly computationally expensive. However these problems were addressed successfully by the development of a new photogrammetry package based upon the DLT. This package would comprise the initial, static, stage of the proposed calibration scheme. An initial investigation was also carried out into the photogrammetric technique known as bundle adjustment which would be required by subsequent stages of the proposed dynamic calibration scheme. A summary of this investigation is presented in Appendix A. However, adequate time was not available to complete the implementation of the full dynamic calibration scheme. A suggested route to its completion is given in section 12.3.

12.2.4. System integration

After the implementation of the static calibration stage was completed, the goals of the project were re-focused. Priority was then given to building a complete demonstration system instead of the continued development of the dynamic calibration scheme. This new goal was accomplished by integrating the ASP sensor system, the MSSM stereo matcher and the DLT based photogrammetry package. The result was a complete, fully automatic, videometrics system, known as CopyCAD. A noteworthy aspect of the system is the provision of a graphical user interface that hides effectively the complexities of the underlying photogrammetry. The ability of the CopyCAD system to capture realistic 3D models of the human face, and hence its potential for use in the chosen application area of maxillo-facial measurement system, was successfully demonstrated.

12.3. Subsequent and future work

The work reported in this thesis has presented many opportunities for continued research in a variety of areas. Some examples will be explored in sub-sections 12.3.1 to 12.3.6. In addition, the work has led directly to the development of a turnkey product for the automatic capture of 3D models and a specialised videometrics system for use in maxillo-facial measurement. These two systems will be described in sub-section 12.3.7.

12.3.1. Analysis of 3D surface reconstruction accuracy

An important stage in the development of any measurement system is the determination of the system's accuracy. The results presented in Chapter 10 indicate

that the developed static calibration procedures provide an acceptable degree of accuracy and that the sensor head provides a suitably stable platform for the stereo cameras. However the accuracy of the complete CopyCAD system, including the MSSM stereo matcher, has not been verified in the field. Unfortunately time was not available to perform the required analysis for the complete system. This is therefore an area that requires further attention, not least to determine whether the original specification for 3D surface measurement accuracy had been met. A simple method for determining the relative 3D measurement accuracy would be to use the CopyCAD system to measure a precisely machined flat metal plate. The 3D measurements obtained from CopyCAD could then be fitted to a plane and the residuals used to quantify the measurement error in the direction normal to the plane.

However a disadvantage of the above approach is that it fails to address the fact that the accuracy of the MSSM stereo matcher, and hence CopyCAD, is highly dependent on the shape of the surface to be measured. A more informative test would therefore be to measure, using the CopyCAD system, a test surface with a shape typical of that expected in a particular application (e.g. a life sized mannequin head in the case of maxillo-facial measurement). The test surface would also have to be measured using a second method, with an accuracy greater than that expected of the CopyCAD system. The two sets of measurements would then have to be transformed to the same co-ordinate frame and subtracted in order to obtain the error in the CopyCAD measurements. Clearly this method would require considerably more work than the flat plane method.

12.3.2. Stereo matcher development

The most significant limitation of the MSSM stereo matcher remaining after the developments reported in this thesis is its poor accuracy in areas of low image texture. This problem was successfully circumvented by using textured scene illumination to avoid such areas occurring. However the provision of textured scene illumination not only adds extra complexity, but is also impractical in many application areas. A desirable improvement to the MSSM stereo matcher would therefore be to incorporate a mechanism to improve surface reconstruction in areas of low image texture. The requirements for accurate shape from shading are complementary to those for accurate shape from stereo, i.e. low amount of image texture and smoothly changing pixel values. Therefore one promising method for improving the accuracy of the MSSM stereo matcher in areas of low image texture would be to add some form of shading constraint.

A second potential improvement to the MSSM stereo matcher would be the explicit integration of camera calibration data. This would allow the epipolar lines in a stereogram to be determined prior to matching and would hence permit the disparity search process to be confined to one dimension, i.e. along the epipolar lines. The result would be a significant reduction in the amount of computation required and hence a corresponding improvement in the speed of the algorithm. Confining the search to the epipolar lines would also increase the match accuracy since the number of potential false matches would be drastically reduced.

12.3.3. Stereo matcher analysis

The investigations carried out into the performance of the MSSM stereo matcher have revealed several areas for continued research. Perhaps the most obvious area is to repeat the performance characterisation experiments that were originally carried using only 1D stereo matching for 2D stereo matching. This would verify that the conclusions drawn from the 1D stereo matching characterisation are indeed also valid for 2D stereo matching.

One significant problem experienced throughout the stereo matcher analysis experiments described in this thesis was that of obtaining suitable test imagery. It was clear that the only practical method of obtaining the large numbers of stereo images with accurately known disparity that were required for the analysis was to synthetically generate them. However, if the synthetic test imagery used in this kind of analysis does not adequately resemble the real imagery that is obtained in practice then it will be very hard to predict the actual level of performance from the results obtained using the synthetic imagery. Although some effort was made in this thesis to address this problem, and a few other researchers have also made limited studies, more work is required to determine effective methods of synthesising truly realistic stereo test imagery.

12.3.4. Static calibration

The only stage in the videometrics process developed in this work that required significant user intervention was the association of world and image target points prior to static calibration. After the completion of the work, Joseph Jin at the Turing Institute developed an automatic method for performing this association. Jin's method relies upon a specially designed calibration object, shown in figure 12.1, in which some of the targets are identified with multiple rings instead of a single target disk. Software based upon the target finder described in section 10.3 was developed to

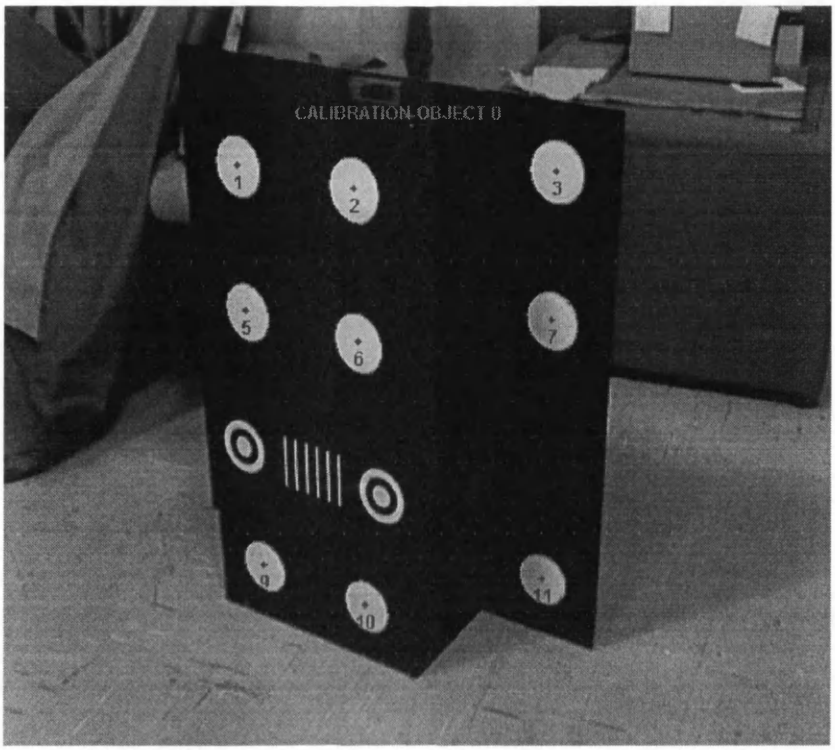


Figure 12.1. New design of calibration object, showing target points that have been detected and identified automatically.

locate automatically these special targets in an image and then identify the remaining targets based upon the locations of the special targets.

12.3.5. Dynamic calibration

Dynamic calibration is the only one of original thesis goals that was not broadly achieved. The purpose of dynamic calibration is to maintain an estimate of the external orientation parameters of a dynamic sensor system's cameras, such as those of the ASP sensor system, when they move. Using the notation developed in Chapter 4, the external orientation parameters can be expressed as ${}^{C1}_{W}\mathbf{T}$ and ${}^{C2}_{W}\mathbf{T}$, the homogeneous transformation matrices representing the transformation from the world co-ordinate system to the Camera 1 and Camera 2 co-ordinate systems respectively. In the case of a dynamic sensor system, these matrices become functions of time, i.e. ${}^{C1}_{W}\mathbf{T}(t)$ and ${}^{C2}_{W}\mathbf{T}(t)$, and the purpose of dynamic calibration is to maintain an estimate of them. Unfortunately, although a full dynamic calibration scheme was proposed, time was not available during the course of the work described in this thesis to complete its implementation. The issues that remain to be addressed in order to complete each of the three stages of the proposed scheme are described below.

12.3.5.1. Static calibration stage

The first stage of the proposed dynamic calibration scheme, static calibration, was intended to calculate the cameras' external orientation parameters for a fixed starting position, i.e. ${}^{C1}_{W}\mathbf{T}(0)$ and ${}^{C2}_{W}\mathbf{T}(0)$, (and also their internal orientation parameters). The DLT based static calibration scheme detailed in Chapter 10 has been shown to be capable of calculating the eleven DLT parameters for a static camera with a fairly high degree of accuracy. Therefore, the DLT based scheme may seem like an ideal candidate for the static stage of dynamic calibration. However, the inter-dependency of the DLT parameters means that they can be used only to obtain an approximate estimate of the actual external (and internal) orientation parameters. Consequently, the DLT based scheme is not sufficient, in itself, to fulfil the requirements of the static stage of dynamic calibration.

In order to address the shortcomings of the DLT, an initial investigation into the more advanced photogrammetric technique of bundle adjustment was carried out by the author, as described in Appendix A. Subsequently, a software package to perform bundle adjustment was developed at the Turing Institute by Tim Niblett. Results obtained using this software have shown that bundle adjustment can successfully refine the approximate external (and internal) orientation parameter values obtained from the DLT and hence achieve very high accuracy. Bundle adjustment can achieve still higher accuracy than the DLT because it can integrate information obtained from multiple images taken with cameras in different positions. The combination of the DLT based calibration scheme described in Chapter 10 (including the refinements described in section 12.3.4), followed by bundle adjustment, is a highly promising method of fulfilling the requirements of the static stage of dynamic calibration.

12.3.5.2. Open-loop stage

The open-loop stage of the dynamic calibration scheme was proposed originally to update the external orientation parameters obtained from static calibration using positional feedback from the ASP sensor head actuators; i.e. to determine ${}^{C1}_{W}\mathbf{T}(t)$ and ${}^{C2}_{W}\mathbf{T}(t)$, the external orientation parameters at time t, given ${}^{C1}_{W}\mathbf{T}(0)$, ${}^{C2}_{W}\mathbf{T}(0)$, the external orientation parameters derived from static calibration, and $a_1(t)$, $a_2(t)$, $e_1(t)$ and $e_2(t)$, the encoder counts for the Camera 1 and Camera 2 azimuth and elevation actuators at time t.

Considering only one of the world to camera transformations, and dropping the camera identification suffix, it was shown in section 4.2 of Chapter 4 that (overleaf):

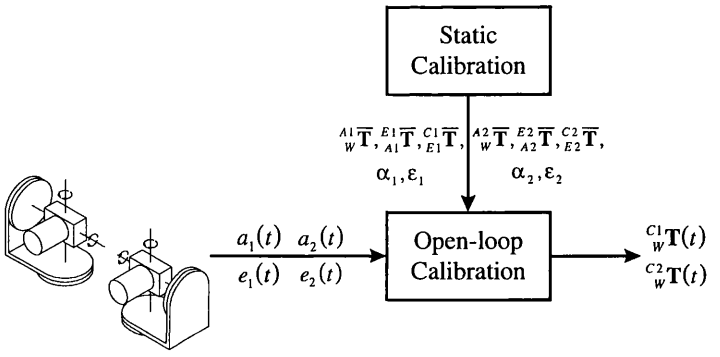


Figure 12.2. Schematic of the open-loop stage of dynamic calibration.

$$ {}^{C}_{W}\mathbf{T} = {}^{C}_{E}\mathbf{T}\,{}^{E}_{A}\mathbf{T}\,{}^{A}_{W}\mathbf{T} \tag{12.1} $$

where ${}^{C}_{W}\mathbf{T}$, ${}^{A}_{W}\mathbf{T}$, ${}^{E}_{A}\mathbf{T}$ and ${}^{C}_{E}\mathbf{T}$ are the homogeneous transformation matrices representing the transformation of co-ordinate systems from, respectively, the world to the camera, the world to the azimuth actuator, the azimuth actuator to the elevation actuator and the elevation actuator to the camera.

In the dynamic case, ${}^{C}_{E}\mathbf{T}$ will be a constant transformation, while both ${}^{A}_{W}\mathbf{T}$ and ${}^{E}_{A}\mathbf{T}$ will contain a constant transformation and a variable transformation that is dependent on the rotation of the azimuth or elevation actuators respectively. The ASP sensor head was designed and constructed so that off-axis wobble in the actuators is negligibly small. Therefore, the variable component of the azimuth actuator transformation can be considered to be a pure rotation about the Y axis, and the variable component of the elevation actuator transformation can be considered to be a pure rotation about the X axis. Equation (12.1) then becomes:

$$ {}^{C}_{W}\mathbf{T}(t) = {}^{C}_{E}\overline{\mathbf{T}}\;\boldsymbol{RotX}\left(\varepsilon \cdot e(t)\right)\,{}^{E}_{A}\overline{\mathbf{T}}\;\boldsymbol{RotY}\left(\alpha \cdot a(t)\right)\,{}^{A}_{W}\overline{\mathbf{T}} \tag{12.2} $$

where, ${}^{A}_{W}\overline{\mathbf{T}}$, ${}^{E}_{A}\overline{\mathbf{T}}$ and ${}^{C}_{E}\overline{\mathbf{T}}$ are constant transformations and ε and α are scale factors required to convert the azimuth and elevation actuator encoder pulses into angles.

The main issue still to be addressed in the implementation of the open-loop stage of dynamic calibration is the development of a method to determine the constant transformations, ${}^{A}_{W}\overline{\mathbf{T}}$, ${}^{E}_{A}\overline{\mathbf{T}}$ and ${}^{C}_{E}\overline{\mathbf{T}}$, and the scale factors, ε and α, for both cameras. One possible solution may be to extend the static calibration stage to step the camera actuators through a large number of positions, acquiring an image at each. The static calibration procedure could then be used to calculate the exterior orientation parameters, i.e. ${}^{C}_{W}\mathbf{T}$, for each position (as well as the set of interior orientation parameters that best fits all of the data). Inverse kinematics could then be applied to

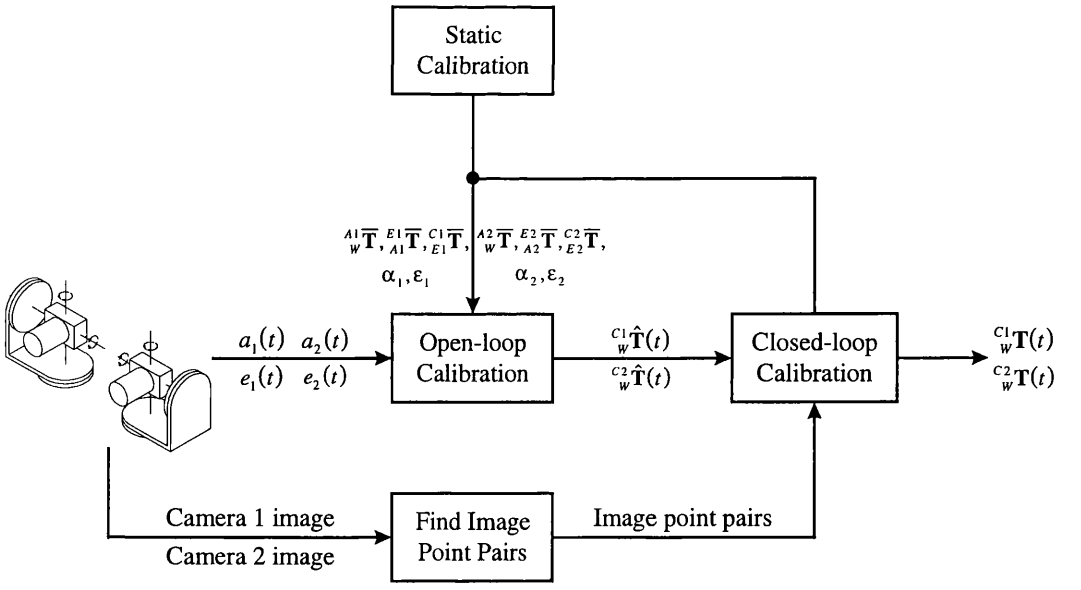


Figure 12.3. Schematic of the closed-loop stage of dynamic calibration.

the set of exterior orientation parameters to obtain the constant transformations ${}^{A}_{W}\overline{\mathbf{T}}$, ${}^{E}_{A}\overline{\mathbf{T}}$ and ${}^{C}_{E}\overline{\mathbf{T}}$, and the constant parameters ε and α. The ASP sensor system may facilitate this process since the rigorous approach used in its design and construction was intended to simplify the fixed transformations, ${}^{A}_{W}\overline{\mathbf{T}}$, ${}^{E}_{A}\overline{\mathbf{T}}$ and ${}^{C}_{E}\overline{\mathbf{T}}$, as much as possible. The resulting open-loop dynamic calibration process is illustrated in figure 12.2.

12.3.5.3. Closed-loop stage

As its name implies, the goal of the closed-loop stage of the dynamic calibration scheme is to close the dynamic calibration loop. It was proposed that this would be done by using information derived from captured images to refine the estimate of the external orientation parameters obtained from open-loop dynamic calibration. It was also proposed that the closed-loop stage may be able to refine the estimates of the internal orientation parameters either in the case of drift or, for example, if the camera lenses were to be zoomed or re-focused.

One possibility for the closed-loop dynamic calibration stage is illustrated in figure 12.3. This scheme would employ bundle adjustment to refine the estimated external orientation parameters derived from a previous stage of open-loop calibration. In addition to the estimates of the exterior orientation parameters, the bundle adjustment would also take as input matched pairs of corresponding image points, as described in Appendix A. It is expected that the accuracy of the initial input provided from open-

loop calibration will allow the bundle adjustment to converge quickly enough for it to be used in real-time.

Clearly, one requirement of the closed-loop stage of dynamic calibration is that the open-loop stage be implemented first. Another requirement is the development of a method to find the matched pairs of corresponding image points in the images acquired from the sensor head. Two methods could be used:

1. Special targets are placed in the scene, e.g. using a calibration object similar to that described in section 12.3.4. This would allow the image point correspondences to be determined automatically, quickly and accurately, using the target detection software already developed. Indeed it might even be possible to modify the target detection software to achieve real time execution rates, in accordance with the expected convergence time of bundle adjustment itself. Using this approach it might then be possible to perform closed-loop calibration during every cycle.

2. The acquired images are matched using the MSSM stereo matcher, and the required number of image point correspondences chosen from the resulting disparity maps. Initially this selection of points could be based purely on the points' match confidence values. However further investigations could be carried out to determine a more optimal selection scheme. The time currently taken for MSSM to match a stereo pair of images precludes this approach from being used during every cycle of a real-time control loop. However it does offer the advantage of not requiring special markers to be placed in the imaged scene.

A final requirement of the closed-loop dynamic calibration is the development of a method to feed the refined exterior orientation parameters back to the open-loop stage of dynamic calibration. This might take the form of corrections to the encoder counts to account for accumulated errors introduced by, for example, back-lash in the actuator gears. Again, inverse kinematics could be used to determine these corrections from the refined external orientation parameters, ${}^{C}_{W}\mathbf{T}$, the known fixed transformations, ${}^{A}_{W}\overline{\mathbf{T}}$, ${}^{E}_{A}\overline{\mathbf{T}}$ and ${}^{C}_{E}\overline{\mathbf{T}}$, and scale factors, ε and α, for each camera.

12.3.6. Dynamic vision

The successful implementation of dynamic calibration would present many opportunities for further work aimed at exploiting the full potential of the ASP system's unique combination of dynamic and quantitative vision. The most obvious example would be the investigation of data fusion from multiple viewpoints to build complete 3D models of large objects. Another example might be the integration of

low-level visual reflexes such as vergence, focus or aperture into the ASP system (as was originally planned). This could allow the system to be used in a semi-autominous manner when direct operator-in-the-loop feedback is difficult, e.g. in tele-operation applications that involve significant information transmission delays. A more ambitious example would be to use the system to provide path planning information for a robot, an autonomous vehicle, or even perhaps a planetary explorer!

12.3.7. Further system development

The CopyCAD system that was developed as part of the work described in this thesis has successfully demonstrated the potential of computer stereo vision based videometrics systems for capturing realistic 3D models. This work has provided the basis for the development, by the Turing Institute, of two new systems: C3D-2020™ and C3D-clinical™. C3D-2020™ is a turn-key product designed for use primarily as a 3D portrait camera. C3D-clinical™ is a specialised, four camera (i.e. two sets of stereo cameras), videometrics system for use in maxillo-facial surgery, which is being developed in collaboration with the Glasgow Dental School and Hospital. The first C3D-2020™ systems are now in use, as shown in figure 12.4 and figure 12.5, and a prototype C3D-clinical™ system is currently undergoing clinical trials at the Glasgow Dental School and Hospital and Canniesburn Hospital, as shown in figure 12.6.

The development of C3D-2020™ and C3D-clinical™ has been based extensively on the results presented in this thesis. For example:

- The process used for the design both systems was based extensively on the methodology developed in this thesis. The ASP sensor system model was also refined and extended in order to aid this design process.

- Both systems use the MSSM stereo matcher. Neither would have been feasible without the development and characterisation of the MSSM stereo matcher presented in this thesis.

- The target location software developed for the CopyCAD system, and improved as described in section 12.3.4, is used in the calibration modules of both systems.

- The photogrammetry used in the static calibration module of CopyCAD and the bundle adjustment software that is described in Appendix A were crucial to the development of the calibration software for both systems.

Further details of C3D-clinical™ can be found in [Siebert, 1994; Mowforth, 1995; Ayoub, 1996].

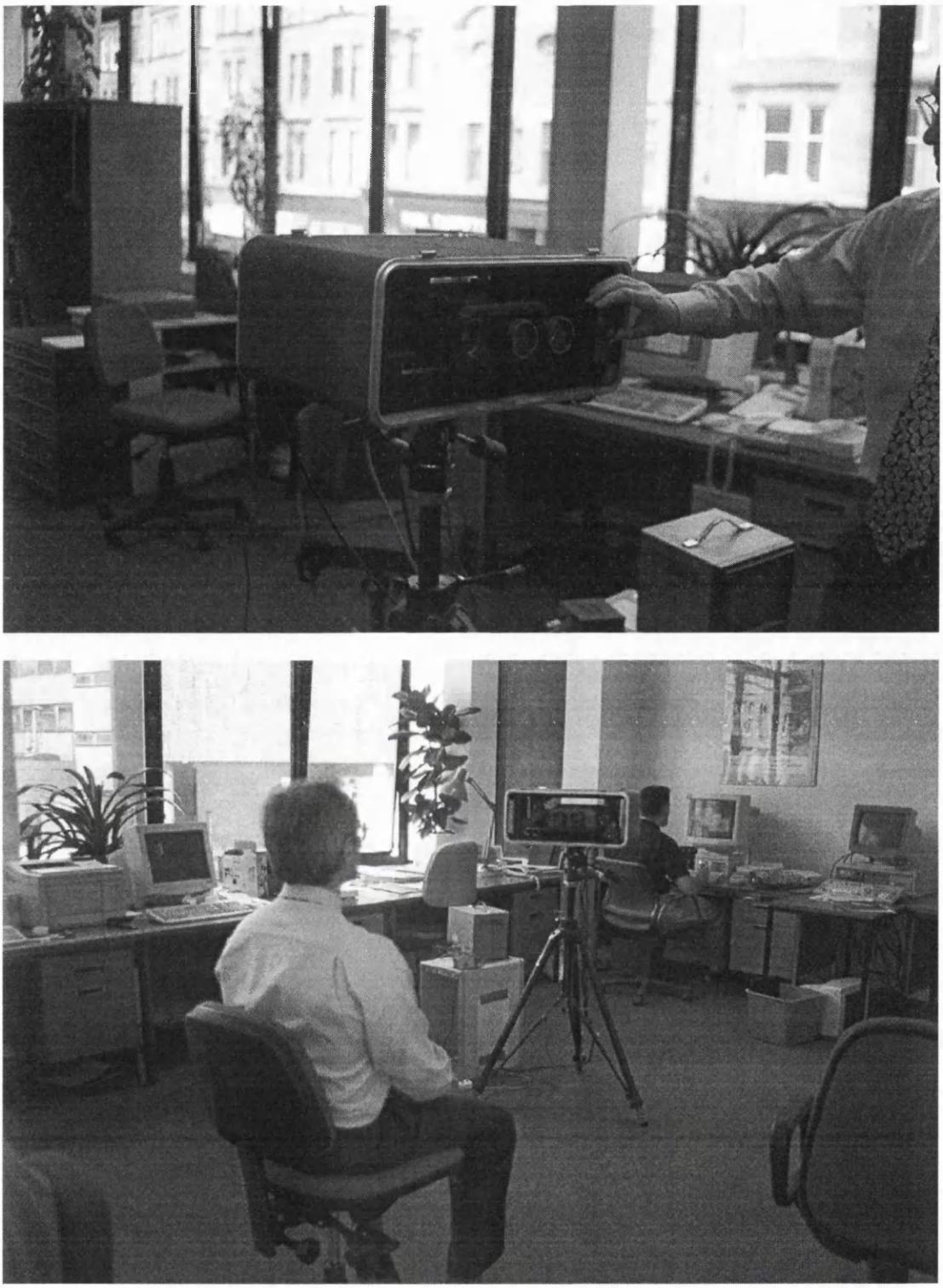


Figure 12.4. C3D-2020™ in use at the Turing Institute.

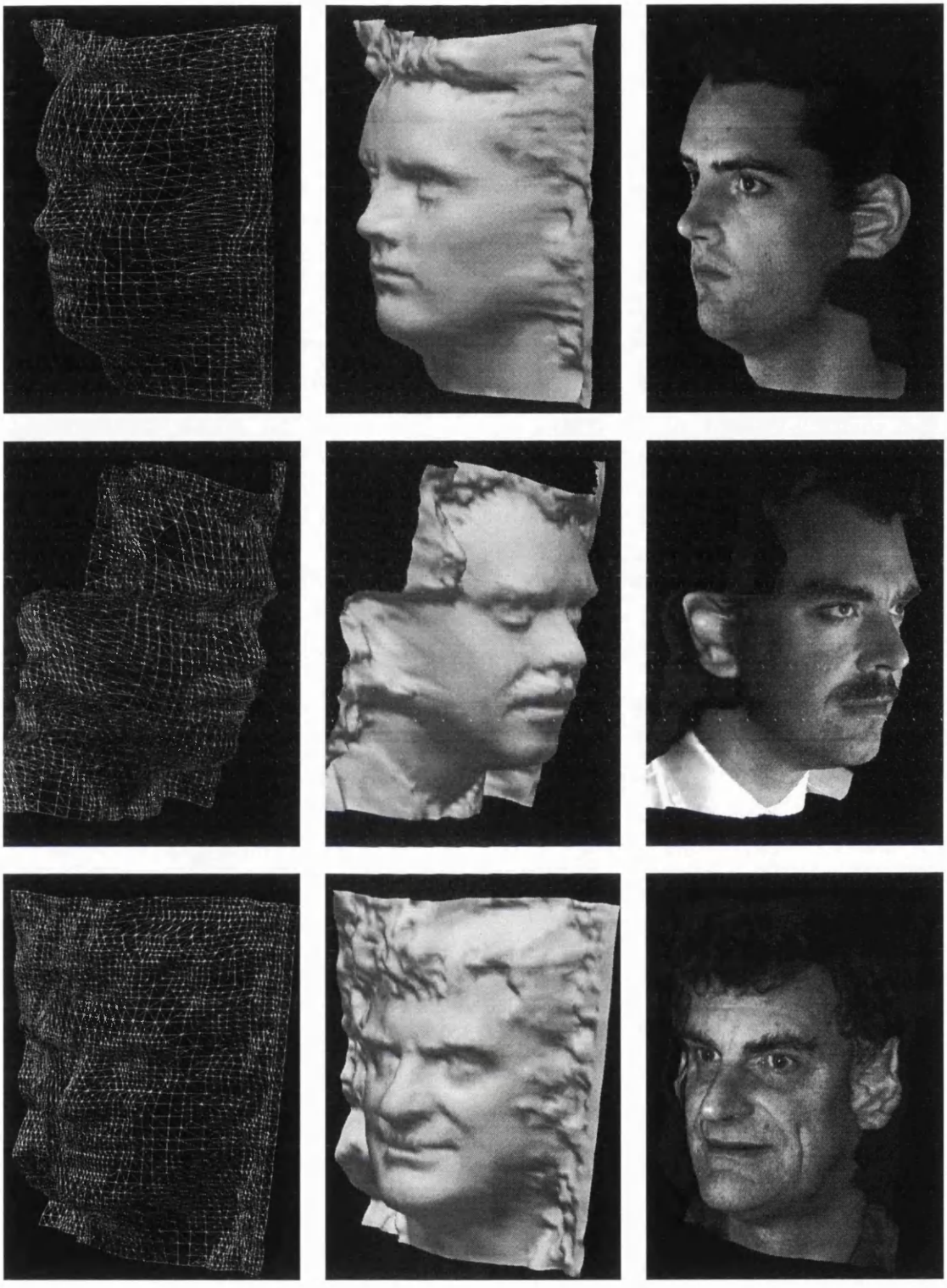


Figure 12.5. Examples of 3D face models captured using a C3D-2020™ system at the 1996 British Machine Vision Conference.

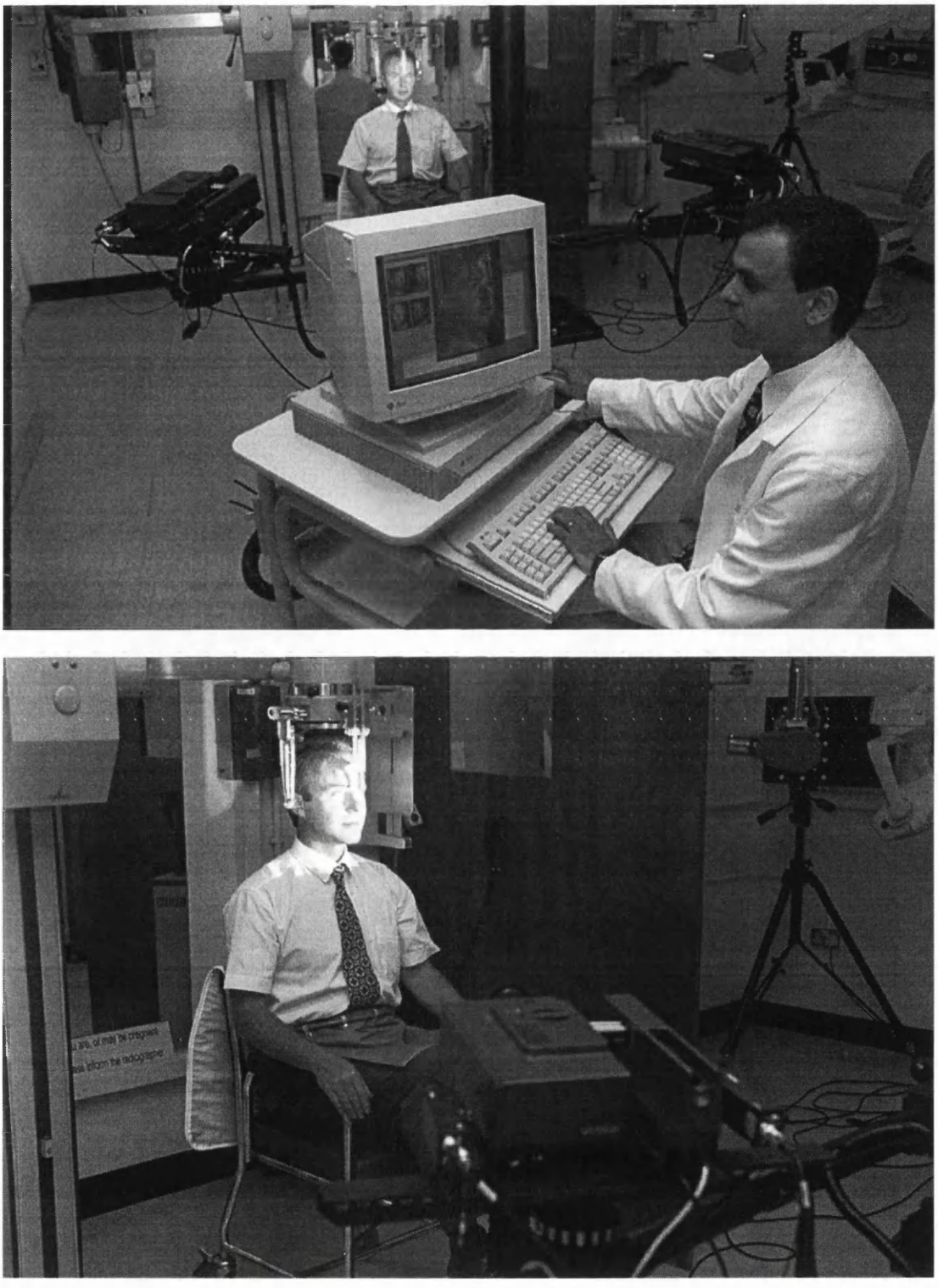


Figure 12.6. Prototype C3D-clinical™ system undergoing trials at Canniesburn Hospital.

12.4. Summary of contributions

The work described in this thesis has made a significant contribution to the field of computer vision in general and videometrics in particular. It has also been successful in integrating research from such diverse fields as computer vision, photogrammetry, mechatronics, human-computer interaction and, more recently, medicine. This has been confirmed by the publication of several refereed academic papers in the proceedings of a variety of national and international conferences. A full list of refereed publications and other presentations of the work described in this thesis is given in the appendices.

The main technical contributions that have been made derive almost directly from the successful achievement of the majority of the original goals of the work. The three main contributions can be summarised as follows:

1. The development of a novel approach to the development of dynamic vision systems based upon the identification of the requirements of a specific visual task early in the design process. This approach makes it possible to specify, design and ultimately fabricate a dynamic sensor "head" that provides the level of performance required for a specific task.

2. The development and analysis of the MSSM stereo matcher which has transformed the algorithm from a laboratory proof-of-principle into a practical and powerful tool.

3. The successful investigation and implementation of photogrammetric techniques and their integration with computer vision to produce a complete videometrics system.

However the vast potential of the work for further exploitation is perhaps equally as important as its academic achievements. Indeed, this potential has already been demonstrated with the successful development of the C3D-2020™ and C3D-clinical™ systems. In addition, the work has contributed significantly to several other commercial projects in fields such as television production, forensic medicine, and facial recognition. It has also been shown that many other opportunities exist to further advance and exploit the work.

References

1. [Abdel-Aziz, 1971] Abdel-Aziz, Y.F. and Karara, N.M., "Direct linear transformation from comparator coordinates into object coordinates in close-range photogrammetry", *Proceedings of the ASP Symposium on Close-Range Photogrammetry*, Illinois, January 1971, pp. 1-18.

2. [Ackermann, 1984] Ackermann, F., "Digital image correlation: performance and potential applications in photogrammetry", *Photogrammetric Record*, Vol. 11, No. 64, October 1984, pp. 429-439.

3. [Aliverti, 1993] Aliverti, A., Ferrigno, G. and Pedotti, A., "Surface analysis by laser beam scanning and stereophotogrammetry", *Videometrics II*, El-Hakim, S.F. (ed.), Proc. SPIE 2067, September 1993, pp. 209-219.

4. [Aloimonos, 1987] Aloimonos, J.Y., Weiss, I. and Bandyopadhyay, A, "Active vision", *Proceedings of the IEEE International Conference on Computer Vision*, London, June 1987, pp. 35-54.

5. [Amin, 1993] Amin, A.M., and Petrie, G., "Geometric calibration of thermal video frame scanners", *Videometrics II*, El-Hakim, S.F. (ed.), Proc. SPIE 2067, September 1993, pp. 44-54.

6. [Anandan, 1987] Anandan, P., *Measuring Visual Motion From Image Sequences,* PhD thesis, Department of Computer and Information Science, University of Massachusetts, May 1987.

7. [Ayache, 1985] Ayache, N. and Faverjon, B., "Fast stereo matching of edge segments using prediction and verification of hypothesis", *Proceedings of Computer Vision and Pattern Recognition,* June 1985, pp. 662-664.

8. [Ayoub, 1996] Ayoub, A.F., Wray, D., Moos, K.F., Siebert, J.P., Jin, J., Niblett, T.B., Urquhart, C.W. and Mowforth, P.H., "three-dimensional modeling for modern diagnosis and planning in maxillofacial surgery", *Journal of Orthognathic and Orthodontic Surgery,* 11, 1996, pp 225-233.

9. [Ballard, 1988] Ballard, D.H., *The Rochester Robot,* University of Rochester Research Report TR-257, Rochester, New York, 1988.

10. [Ballard, 1991] Ballard, D.H., "Animate vision", *Artificial Intelligence,* Vol. 48, No. 1, February 1991, pp. 57-86.

11. [Bajcsy, 1985] Bajcsy, R., "Active perception vs. passive perception", *Proceedings of the 3rd Workshop on Computer Vision,* Michigan, USA, October 1985, pp. 55-59.

12. [Bani-Hashemi, 1991] Bani-Hashemi, A., "Finding the aspect ratio of an imaging system", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 1991, pp. 122-126.

13. [Bernard, M.] Bernard, M., Boutaleb, A.K., Kölbl, O. and Penis, C., "Automatic stereophotogrammetry: implementation and comparison of classical corelation methods and dynamic programming based techniques", *Proceedings of ISPRS Commision III International Symposium From Analytical to Digital,* Rovaniemi, Finland, 1986.

14. [Bernard, 1980] Bernard, S.T. and Thomson, W.B., "Disparity analysis of images", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. PAMI-2, No. 4, July 1980, pp. 333-340.

15. [Bernard, 1990] Bernard, S.T. and Fischler, M.A., "Computational and biological models of stereo vision", *Proceedings of DARPA Image Understanding Workshop*, Pittsburgh, Penssylvannia, September 1990, pp. 439-448.

16. [Besl, 1988] Besl, P.J., "Active optical range imaging sensors", *Machine Vision and Applications*, Vol. 1, No. 2, 1988, pp. 127-152.

17. [Beyer, 1992] Beyer, H.A., *Geometric and radiometric analysis of a CCD-camera based photogrammetric close-range system*, Institute of Geodesy and Photogrammetry, Zurich, May 1992.

18. [Bopp, 1977] Bopp, H. and Krauss, H., "A simple and rapidly converging orientation and calibration method for non-topographic applications", *Proceedings of the ASP Fall Meeting*, October 1977.

19. [Brandow, 1976] Brandow, V.D., Karara, H.M., Damberger, H.H. and Krausse, H.F., *Photogrammetric Engineering and Remote Sensing*, Vol. 42, No. 5, May 1976, pp. 637-648.

20. [Brown, 1971] Brown, D.C., "Close-range camera calibration", *Photogrammetric Engineering*, Vol. 37, No. 8, 1971, pp. 855-866.

21. [Burner, 1985] Burner A.W., Snow, W.L. and Goad, W.K., "Close-range photogrammetry with video cameras", *Technical Papers of the 51st ASP Anual Meeting*, Vol 1, Washington D.C., March 10-15, 1985.

22. [Burt, 1983] Burt, P. and Adelson, E, "The laplacian pyramid as a compact image code", *IEEE Transactions on Communications*, Vol. COM-31, No. 4, 1983, pp. 532-540.

23. [Burt, 1988] Burt, P., "Algoritms and architectures for smart sensing", *Proceedings of the DARPA Image Understanding Workshop*, Cambridge MA, Vol. 1, 1988, pp. 139-153.

24. [Buxton, 1991] Buxton, B., "European heads", *BMVA News*, Hancock, E. (ed.), Vol. 2, No. 3, August 1991.

25. [Canny, 1986] Canny, J.F., "A computational approach to edge detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 8, No. 6, November 1986, pp. 679-698.

26. [Caprile, 1990] Caprile, B. and Torre, V., "Using vanishing points for camera calibration", *International Journal of Computer Vision,* No. 4, 1990, pp. 127-140

27. [Church, 1945] Church, E., *Revised Geometry of the Aerial Photograph - Bulletin of Aerial Photogrammetry*, No. 15, Syracuse University, 1945.

28. [Cochran, 1989]Cochran, S. and Medioni, G., "Accurate surface description from binocular stereo", *Proceedings of the DARPA Image Understanding Workshop*, Palo Alto CA, May 1989, pp. 857-869.

29. [Cohen, 1991] Cohen, Y, Landy, M.S., *The HIPS Image Processing Software: HIPS-2 Version*, SharpImage Software, New York, August 1991.

30. [Cumani, 1991] Cumani, A., Guiducci, A., Grattoni, P., Pettiti, G. and Pallastri, F., "High accuracy localisation of calibration points for dimensional

measurements by image processing techniques", *Proceedings of the Fifth International Conference on Advanced Robotics*, Pisa, Italy, 18-22 June 1991, pp. 1761-1765.

31. [Curry, 1985] Curry, S., Baumrind, S. and Anderson, J.M., "Calibration of an array camera", *Technical Papers of the 51st ASP Annual Meeting,* Vol. 1, Washington DC., March 1985, pp. 331-340.

32. [Cyberware, 1995] Cyberware, *Cyberware Whole Body Scanning*, Product information, Cyberware, 2110 Del Monte Avenue, Monterey, California 93940, USA.

33. [DataCell, 1992a] DataCell Limted, *S2200 Real Time Colour Frame Grabber and Display User Manual*, Version 1.1, Reading, UK, 1992.

34. [DataCell, 1992b] DataCell Limted, *ITEX S2200 Programmer's Manual*, Reading, UK, 1992.

35. [Das, 1990] Das, S. and Ahuja, N., "Active stereo based surface reconstruction", *Proceedings of the 5th IEEE International Symposium on Intelligent Control,* Philadelphia PA., September 1990, pp. 227-232.

36. [Day 1989] Day, T. and Muller J.P., "Digital elevation model production by stereo-matching SPOT image-pairs: a comparison of algorithms", *Image and Vision Computing*, Vol. 7, No.2, May 1989, pp. 95-101.

37. [Dhond, 1989] Dhond, U.R. and Aggarwal, J.K., "Structure from stereo - a review", *IEEE Transactions on Systems, Man, and Cybernetics,* Vol. 19, No. 6, December 1989.

38. [Driels, 1991] Driels, M.R. and Pathre, U.S., "Vision-based automatic theodolite for robot calibration", *IEEE Transactions on Robotics and Automation*, Vol. 7, No. 3, June 1991, pp. 351-360.

39. [Ealing, 1990] Ealing Electro-Optics plc., *Product Catalogue*, 1990.

40. [Echigo] Echigo, T., "A camera calibration technique using three sets of parallel lines", *Machine Vision and Applications*, No. 3, 1990, pp. 159-167.

41. [El-Hakim] El-Hakim, S.F., "Introduction", *Videometrics II*, El-Hakim, S.F. (ed.), Proc. SPIE 2067, September 1993, pp. vii.

42. [Faig, 1975] Faig, W., "Calibration of close-range photogrammetry system: mathematical formulation", *Photogrammetric Engineering and Remote Sensing*, Vol. 41, No. 12, December 1975, pp. 1479-1486.

43. [Faig, 1976] Faig, W., "Photogrammetric potential of non-metric cameras", *Photogrammetric Engineering and Remote Sensing*, Vol. 42, No. 1, January 1976, pp. 47-49.

44. [Fan, 1993] Fan, H. and Baozong, Y., "High performance camera calibration algorithm", *Videometrics II*, El-Hakim, S.F. (ed.), Proc. SPIE 2067, September 1993, pp. 2-13.

45. [Faugeras, 1986] Faugeras, O.D. and Toscani, G., "The calibration problem for stereo", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* Miami Beach, Florida, June 1986, pp. 15-20.

46. [Granshaw, 1980] Granshaw, S.I., "Bundle adjustment methods in engineering photogrammetry", *Photogrammetric Record*, Vol. 10, No. 56, 1980, pp. 181-207.

47. [Grattoni, 1990] Grattoni, P. and Guiducci, P., "Contour coding for image desciption", *Pattern Recognition Letters*, 11, February 1990, pp. 95-105.

48. [Grimson, 1985] Grimson, W.E.L., "Computational experiments with a feature based stereo algorithm", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 7, No. 1, January 1985, pp. 17-34.

49. [Gruen, 1988] Gruen, A.W. and Baltsavias, E., "Automatic 3-D measurement of human faces with CCD cameras", *SPIE Vol. 1030*, 1988, pp. 106-116.

50. [Gruen, 1992] Gruen, A., "Recent advances of photogrammetry in robot vision", *ISPRS Journal of Photogrammetry and Remote Sensing*, No. 47, 1992, pp. 307-323.

51. [Gülch, 1984] Gülch, E., "Calibration of two CCD-cameras used for digital image correlation on the planicomp C100", *International Archives of Photogrammetry and Remote Sensing,* Vol. XXV, Part A3a, 1984, pp. 159-168.

52. [Haleva, 1957] Haleva, U.V., "", *Proceedings of the International Photogrammetric Conference on Aerial Triangulation*, Ottawa, Canada, August 1957.

53. [Hertzberg, 1957] Hertzberg, H.T.E., Dupertuis, H.T.E. and Emmanuel, I., "Stereophotogrammetry as an anthropometric tool" *Photogrammetric Engineering*, No. 23, 1957, pp. 942-947.

54. [Hoff, 1989] Hoff, W. and Ahuja, N., "Surfaces from stereo: integrated feature matching, disparity estimation, and contour detection", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 12, No. 2, February 1989, pp. 121-136.

55. [Jansa, 1993] Jansa, J., Huang, Y. and Trinder, J.C., "Problems of precise target location and camera orientation in digital close-range photogrammetry", *Videometrics II*, El-Hakim, S.F. (ed.), Proc. SPIE 2067, September 1993, pp. 151-161.

56. [Jarvis, 1983] Jarvis, R.A., "A perspective on range finding techniques for computer vision", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 5, No. 2, March 1983, pp. 122-139.

57. [Johannesson, 1993] Johannesson, M., *Sheet-of-Light Range Imaging*, PhD theis, No. 404, LIU-TEK-LIC-1993:46, Linköping University, Sweden, November 1993.

58. [Karara, 1974] Karara, H.M. and Abdel-Aziz, Y.I., "Accuracy aspects of non-metric imageries", *Photogrammetric Engineering*, September 1974.

59. [Karara, 1989] Karara, H.M. (ed.), *Handbook of Non-Topographic Photogrammetry*, 2nd edition, American Society for Photogrammetry and Remote Sensing, Falls Church, VA., 1989.

60. [Kölbl, 1976] Kölbl, O.R., "Metric or non-metric cameras", *Photogrammetric Engineering and Remote Sensing*, Vol. 42, No. 1, January 1976, pp. 103-113.

61. [Kratky, 1979] Kratky, V., "Real-time photogrammetric support of dynamic three-dimensional control", *Photogrammetric Engineering and Remote Sensing*, Vol. 45, 1979, pp. 1231-1242.

62. [Krotkov, 1989] Krotkov, E.P., *Active computer vision by cooperative focus and stereo*, Springer-Verlag, New York, 1989.

63. [Landy, 1984] Landy, M.S., Cohen, Y. and Sperling, G., "HIPS: a unix based image processing system", *Computer Vision, Graphics and Image Processing*, Vol. 25, 1984, pp. 331-347.

64. [Landy, 1991] Landy, M.S., *A Programmers Guide to the HIPS Software: HIPS2 Version*, SharpImage Software, New York, Januray 1991.

65. [Lenz, 1988] Lenz, R.K. and Tsai, R.Y, "Techniques for calibrating the scale factor and image center for high accuracy 3-D machine vision metrology" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, No. 5, September 1989, pp. 713-720.

66. [McDonald, 1993] McDonald, J.P., Siebert, J.P. and Fryer, R.J., "Stereo scene coding using SLM active illumination", *Proceedings of the 26th International Symposium on Automotive Technology and Automation, Mechatronics Conference,* Aachen, Germany, September 1993, pp. 169-176.

67. [McDonald, 1994] McDonald, J.P., *Developments in Active Ranging Using an Advanced Stereo/Projection Test Bed*, PhD thesis, The University of Strathclyde, 1994.

68. [Marr, 1982] Marr, D., *Vision*, W.H. Freeman and Company, New York, 1982.

69. [Marr, 1979] Marr, D. and Poggio, T., "A computation theory of human stereo vision", *Proceedings of the Royal Society of London*, Vol. B207, 1979, pp. 187-217.

70. [Marshall, 1991] Marshall, S.J., *et al.*, "Data capture techniques for 3-D facial imaging", *Computer Vision and Image Processing,* Barret, A. (ed.), Chapman and Hall, 1991, pp. 248-275.

71. [Maybank, 1992] Maybank, S.J. and Faugeras, O.D., "A theory of self-calibration of a moving camera", *International Journal of Computer Vision,* Vol. 8, No. 2, 1992, pp. 123-151.

72. [Medioni, 1985]Medioni, G. and Nevatia, R., "Segment based stereo matching", *Computer Vision, Graphics, and Image Processing*, Vol. 31, July 1985, pp. 2-18.

73. [Micro-Controle, 1990] Micro-Conreole Ltd., *Product Catalogue*, 1990.

74. [Microsoft, 1994] The Microsoft Corporation, *Microsoft Excel User's Guide.* Version 5.0, 1994, pp. 599-600

75. [Mowforth, 1995] Mowforth, P., Ayoub, A., Jin, J., Moos, K., Niblett, T., Siebert, P., Urquhart, C. and Wray, D., "3D imaging system for clinical applications", *Medical Electronics*, December 1995, pp. 59-63.

76. [Mohan, 1989] Mohan, R., Medioni, G. and Nevatia, R., "Stereo error detection, correction and evaluation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 11, No. 2, February 1989, pp. 113-120.

77. [Nishihara, 1987] Nishihara, H.K., "Practical real-time imaging stereo matcher", *Readings in Computer Vision: Issues, Problems, Principles and Paradigms*, Fischeler, M.A. and Firschein, O. (eds.), Morgan Kaufmann, Los Altos CA, pp. 83-72.

78. [Nasu, 1976] Nasu, M., "Statistical testing procedures applied to analytical camera calibration of non-metric systems", *Photogrammetric Engineering and Remote Sensing*, Vol. 42, No. 6, June 1976, pp. 777-788.

79. [Ohta, 1985] Ohta, Y. and Kanade, T., "Stereo by intra- and inter-scanline search using dynamic programming", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 7, No. 2, February 1985, pp. 139-154.

80. [Ousterhaut, 1994] Ousterhaut, J.K., *Tcl and the Tk Toolkit*, Addison-Wesley Publishing Company, Reading, Massachusetts, April, 1994.

81. [Pahlavan, 1991] Pahlavan, K. and Eklundh, J.O., *A Head-Eye System for Active Purposive Computer Vision*, Technical Report CVAP-80, Department of Numerical Analysis and Computer Science, Royal Institute of Technology, Stockholm, 1991.

82. [Penna, 1991] Penna, M.A., "Camera calibration: a quick and easy way to determine the scale factor", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 12, December 1991, pp. 1240-1245.

83. [Physik Instrumente, 1990] Physik Instrumente GmbH. & Co., *Product Catalogue*, 1990.

84. [Pinkney, 1978] Pinkney, H.F.L., "Theory and development of an on-line 30Hz video photogrammetric system for real-time three-dimensional control", *Proceedings of ISP Symposium on Photogrammetry for Industry*, August 1978, Stockholm.

85. [Prager, 1983] Prager, J.M. and Arbib, M.A., "Computing the optic flow: the MATCH algorithm and prediction", *Computer Vision, Graphics, and Image Processing*, No. 24, 1983, pp. 271-304.

86. [Robson, 1993] Robson, S., Clarke, T.A. and Chen, J., "The suitability of the Pulnix TM6CN CCD camera for photogrammetric measurement", *Proceedings of SPIE Videometrics II*, Proc. SPIE 2067, September 1193, pp. 66-77.

87. [Sato, 1987] Sato, K. and Inokuchi, S., "Range-imaging system utilizing nematic liquid crystal mask", *Proceedings of the IEEE International Conference on Computer Vision*, London, June 1987, pp. 657-661.

88. [Siebert, 1990] Siebert, J.P. and Urquhart, C.W., "Active stereo: texture enhanced reconstruction", *Electronics Letters*, **26** (7), March 1990.

89. [Siebert 1994] Siebert, J.P. and Urquhart, C.W., "C3D: a novel vision-based 3-D data acquisition system", *Proceedings of the European Workshop on Combined Real and Synthetic Image Processing for Broadcast Video Production*, Hamburg, Germany, November 1994, Springer Verlag, pp. 170-180.

90. [Skelly, 1992] Skelly, I.J., *Medical/CAD modelling using vision,* BSc/BEng Final Year Report, University of Strathclyde, UK., April 1992.

91. [Slama, 1980] Slama, C.C. (ed.), *Manual of photogrammetry*, 4th edition, American Society of Photogrammetry, Falls Church, VA., 1980.

92. [Sommerville, 1951] Sommerville, D.M.Y., *Analytical Geometry of Three Dimensions,* Cambridge University Press, Cambridge, U.K., 1951.

93. [Sony, 1990] The Sony Corporation, *Sony CCD Video Camera Module (XC77RR-CE) Service Manual*, 1990.

94. [Thacker, 1992] Thacker, N.A. and Courtney, P., *Calibration of symmetric and asymmetric control paradigms for a 4 DOF head*, Artificial Intelligence Vision Research Unit, University of Shefield, AIVRU ref. no. 65, January 1992.

95. [Thompson, 1966] Thompson, M.M. (ed.), *Manual of photogrammetry*, 3rd edition, American Society of Photogrammetry, Falls Church, VA., 1966.

96. [Toscani, 1989] Toscani, G., Vaillant, R., Deriche, R. and Faugeras, O.D., "Stereo camera calibration using the environment", *Proceedings of the 6th Scandanavian Conference on Image Analysis*, Oulu, Finland, June 1989, pp. 953-960.

97. [Trucco, 1994] Trucco, E., Fisher, R.B. and Fitzgibbon, A.W., "Direct calibration and data consistency in 3-D laser scanning", *Proceedings of the British Machine Vision Conference*, September 1994, pp. 489-498.

98. [Tsai, 1986] Tsai, R.Y., "An efficient and accurate camera calibration technique for 3D machine vision", *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, Miami Beach, Florida, June 1986, pp. 364-374.

99. [Tsai, 1987] Tsai, R.Y., "A versitile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses", *IEEE Journal of Robotics and Automation*, Vol. 3, No. 4, August 1987, pp. 323-344.

100. [Tsai, 1989] Tsai, R.Y., "Synopsis of recent progress on camera calibration for 3D machine vision", *The Robotics Review,* Khatib, O., Craig, J.J. and Lozano-Perez, T. (eds.), MIT Press, 1989, pp. 147-159.

101. [Undbekken, 1991] Undbekken, K., Wilson, D., Mowforth, P., Solbakken, S. and Eiklid, B., "Design of an anthropomorphic robot head", *Proceedings of the British Machine Vision Conference*, Glasgow, Scotland, September 1991, pp. 387-391.

102. [Urquhart, 1989] Urquhart, C.W., *Automated methods for the extraction of depth information from stereograms*, industrial placement report, BBN Systems and Technologies and Heriot-Watt University, September 1989.

103. [Urquhart, 1990] Urquhart, C.W., *An investigation into active and passive methods of improving the performance of scale-space stereo*, MEng thesis, BBN Systems and Technologies and Heriot-Watt University, April 1990.

104. [Van Hoff, 1992a] Van Hoff, A.A., *Efficient computation of Gaussian pyramids*, Research report, The Turing Institute, August 1992.

105. [Van Hoff, 1992b] Van Hoff, A.A., *An efficient implementation of MSSM*, Research report, The Turing Institute, December 1992.

106. [Van Wijk, 1976] Van Wijk, M.C. and Ziemann, H., "The use of non-metric cameras in monitoring high speed processes", *Photogrammetric Engineering and Remote Sensing*, Vol. 42, No. 1, January 1976, pp. 91-102.

107. [Wang, 1991] Wang, L., and Tsai, W., "Camera calibration by vanishing lines for 3D computer vision", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 13, No. 4, April 1991, pp. 370-376.

108. [Weng, 1990] Weng, J., Cohen, P and Herniou, M., "Calibration of stereo cameras using a non-linear distortion model", *Procedings of the 10th IEEE International Conference on Pattern Recognition*, 1990, pp. 246-253.

109. [Weng, 1992] Weng, J., Cohen, P and Herniou, M., "Camera calibration with distortion models and accuracy evaluation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 10, October 1992, pp. 965-980.

110. [Weiss, 1991] Weiss, N.A. and Hassett, M.J., *Introductory Statistics*, 3rd Edition, Addison-Wesley Publishing Company Inc., USA, 1991, pp. 596 - 632.

111. [Wong, 1975] "Mathematical formulation and digital analysis in close-range photogrammetry", *Photogrammetric Engineering and Remote Sensing*, Vol. 41, No. 11, November 1975, pp. 1355-1373.

112. [Yau, 1988] Yau, J.F.S. and Duffy, N.D., "A texture mapping approach to 3D facial image synthesis", *Computer Graphics Forum 7,* 1988, pp. 129-134.

113. [Yau, 1989] Yau, J.F.S., *A model based approach to picture phone coding*, PhD thesis, Heriot-Watt University, Edinburgh, UK., 1989.

114. [Zhang, 1994] Zhang, Z., Luong, Q.T. and Faugeras, O.D., "Self-calibration of an uncalibrated stereo rig from one unknown motion", *Proccedings of the British Machine Vision Conference,* September 1994, pp. 499-508.

115. [Zhengping, 1988] Zhengping, J., *On the multi-scale iconic representation for low-level computer vision systems*, PhD thesis, The Turing Institute and The University of Strathclyde, 1988.

116. [Zhengping, 1989] Zhegping, J. and Mowforth, P., *A discrete approach to signal matching*, Research Memo TIRM-89-036, The Turing Institute, Glasgow, Scotland, 1989.

Publications

1. Siebert, J.P. and Urquhart, C.W., "Active stereo: texture enhanced reconstruction", *Electronics Letters*, IEE, Vol 26, No. 7, March 1990, pp. 427-430.

2. Mowforth, P.H., Siebert, J.P., Zhengping, J. and Urquhart, C.W., "A head called Richard", *Proceedings of the British Machine Vision Conference*, University of Oxford, U.K., Spetember 1990, pp. 361-365.

3. Siebert, J.P., Urquhart, C.W., Wilson, D.F., McDonald, J.P., Mowforth, P.H. and Fryer, R.J., "The active stereo probe: dynamic video feedback", *Proceedings of the British Machine Vision Conference*, University of Glasgow, September 1991, pp. 381-386.

4. Urquhart, C.W. and Siebert, J.P., "Development of a precision active stereo system", *Proceedings of the IEEE International Symposium on Intelligent Control*, Glasgow, Scotland, August 1992, pp. 354-359.

5. Urquhart, C.W., McDonald, J.P., Siebert, J.P. and Fryer, R.J., "Active animate stereo vision", *Proceedings of the British Machine Vision Conference*, University of Surrey, U.K., September 1993, pp. 75-84.

6. Urquhart, C.W. and Siebert, J.P., "Towards real-time dynamic close range photogrammetry", *Videometrics II*, El-Hakim, S.F. (ed.), Proc. SPIE 2067, September 1993, pp. 240-251.

7. McDonald, J.P., Siebert, J.P., Fryer, R.J. and Urquhart, C.W., "Visualisation and model building in medical imaging", *Medical Informatics*, Vol. 19, No. 1, 1994, pp. 61-69.

8. Siebert, J.P. and Urquhart, C.W., "C3D: a novel vision-based 3-D data acquisition system", *Proceedings of the European Workshop on Combined Real and Synthetic Image Processing for Broadcast Video Production*, Hamburg, Germany, November 1994, Springer Verlag, pp. 170-180.

9. Mowforth, P.H., Ayoub, A.F., Jin, J., Moos, K.F., Niblett, T.B., Siebert, J.P., Urquhart, C.W. and Wray, D., "3D imaging system for clinical applications", *Medical Electronics*, December 1995, pp. 59-63.

10. Ayoub, A.F., Wray, D., Moos, K.F., Siebert, J.P., Jin, J., Niblett, T.B., Urquhart, C.W. and Mowforth, P.H., "Three-dimensional modelling for modern diagnosis and planning in maxillofacial surgery", *Journal of Adult Orthognathic and Orthodontic Surgery,* 11, 1996, pp 225-233.

Appendix A
Bundle adjustment

A.1. Introduction

Bundle adjustment is a very powerful and flexible method of performing space resection that was first developed by photogrammetrists in the early 1970s, e.g. [Brown, 1971; Faig, 1975; Wong, 1975; Granshaw, 1980]. The basic problem solved by bundle adjustment is to determine the interior and exterior orientation parameters of a number of cameras, given the co-ordinates of corresponding image points in each of the cameras. This is done by first forming a mathematical model in which each of the cameras is represented using the collinearity equations (see equations (2.6) to (2.14) in Chapter 2). When an image point is added into this collinearity equation based model, it is represented by a straight ray of light that passes through the image point itself, the perspective centre of the camera and the (possibly unknown) corresponding world point. The result of adding all of the observed image points into the model is a "bundle" of such rays of light passing through the perspective centre of each camera. The problem is then to iteratively adjust the parameters of each camera, and hence its bundle of rays, until all of the rays of light from different cameras that correspond to the same world point intersect as closely as possible at a single point in space.

This appendix describes a preliminary investigation into bundle adjustment that was carried out by the author. Firstly, a mathematical model of bundle adjustment is derived in section A.2. This is based extensively on the derivation given in [Slama, 1980]. Secondly, in section A.3, a description is given of a computer program to perform bundle adjustment that was implemented by the author. Unfortunately, sufficient time was not available to complete the testing of this computer program. Therefore, it is not possible to present any results of the use of the program.

A.2. Derivation

A.2.1. Problem formulation

The basic formulation of bundle adjustment assumes that n world points, $\mathbf{P}_j, j = 1,2,\cdots,n$, are imaged by m cameras, $\mathbf{C}_i, i = 1,2,\cdots,m$. The result is $m \times n$ image points, $\mathbf{p}_{ij}, i = 1,2,\cdots,m, j = 1,2,\cdots,n$:

$$\mathbf{C}_i = \begin{bmatrix} \omega_i & \phi_i & \kappa_i & X_{io} & Y_{io} & Z_{io} & x_{io} & y_{io} & c_i \end{bmatrix}^{\mathrm{T}}, i = 1,2,\cdots,m \tag{A.1}$$

$$\mathbf{P}_j = \begin{bmatrix} X_j & Y_j & Z_j \end{bmatrix}^{\mathrm{T}}, j = 1,2,\cdots n \tag{A.2}$$

$$\mathbf{p}_{ij} = \begin{bmatrix} x_{ij} & y_{ij} \end{bmatrix}^{\mathrm{T}}, i = 1,2,\cdots,m, \; j = 1,2,\cdots,n \tag{A.3}$$

where (X_{i0}, Y_{i0}, Z_{i0}) are the world co-ordinates of the perspective centre of camera $\mathbf{C}_i$, ω_i, ϕ_i and κ_i are the three Euler angles defining the orientation of camera $\mathbf{C}_i$, (x_{i0}, y_{i0}) are the image co-ordinates of the principal point for camera $\mathbf{C}_i$, c_i is the principal distance of camera $\mathbf{C}_i$, (X_j, Y_j, Z_j) are the co-ordinates of world point $\mathbf{P}_j$; and (x_{ij}, y_{ij}) are the co-ordinates of image point $\mathbf{p}_{ij}$, which is world point $\mathbf{P}_j$ imaged by camera $\mathbf{C}_i$.

The collinearity equations (as defined in equations (2.18) and (2.19) in Chapter 2) for each image point, $\mathbf{p}_{ij}$, can be expressed as two functions, $f_{x_{ij}}$ and $f_{y_{ij}}$:

$$f_{x_{ij}}\left(\mathbf{C}_i, \mathbf{P}_j, \mathbf{p}_{ij}\right) = c_i \frac{r_{i11}(X_j - X_0) + r_{i12}(Y_j - Y_{i0}) + r_{i13}(Z_j - Z_{i0})}{r_{i31}(X_j - X_{i0}) + r_{i32}(Y_j - Y_{i0}) + r_{i33}(Z_j - Z_{i0})} - x_{ij} + x_{i0} = 0 \tag{A.4}$$

$$f_{y_{ij}}\left(\mathbf{C}_i, \mathbf{P}_j, \mathbf{p}_{ij}\right) = c_i \frac{r_{i21}(X_j - X_{i0}) + r_{i22}(Y_j - Y_{i0}) + r_{i23}(Z_j - Z_{i0})}{r_{i31}(X_j - X_{i0}) + r_{i32}(Y_j - Y_{i0}) + r_{i33}(Z_j - Z_{i0})} - y_{ij} + y_{i0} = 0 \tag{A.5}$$

where $r_{i11}, \cdots, r_{i33}$ are the elements of the rotation matrix for camera $\mathbf{C}_i$ formed from ω_i, ϕ_i and κ_i, as shown in equations (2.6) to (2.14) in Chapter 2.

The general problem addressed by bundle adjustment is to solve equations (A.4) and (A.5) for any or all of the other parameters given measurements of each image point $\mathbf{p}_{ij}$:

$$\begin{bmatrix} \hat{x}_{ij} \\ \hat{y}_{ij} \end{bmatrix} = \begin{bmatrix} x_{ij} \\ y_{ij} \end{bmatrix} - \begin{bmatrix} v_{x_{ij}} \\ v_{y_{ij}} \end{bmatrix} \tag{A.6}$$

$$\hat{\mathbf{p}}_{ij} = \mathbf{p}_{ij} - \mathbf{v}_{ij} \tag{A.7}$$

where (x_{ij}, y_{ij}) are the (unknown) actual co-ordinates of image point $\mathbf{p}_{ij}$, $(\hat{x}_{ij}, \hat{y}_{ij})$ are the measured or observed co-ordinates of image point $\mathbf{p}_{ij}$, and $(v_{x_{ij}}, v_{y_{ij}})$ are the (again unknown) residuals between the actual and measured co-ordinates.

In order to solve equations (A.4) and (A.5), it is first necessary to linearise them. Linearisation requires that initial approximations to the camera parameters and world co-ordinates be available:

$$\begin{bmatrix} \omega_i \\ \phi_i \\ \kappa_i \\ X_{i0} \\ Y_{i0} \\ Z_{i0} \\ x_{i0} \\ y_{i0} \\ c_i \end{bmatrix} = \begin{bmatrix} \omega_i^0 \\ \phi_i^0 \\ \kappa_i^0 \\ X_{i0}^0 \\ Y_{i0}^0 \\ Z_{i0}^0 \\ x_{i0}^0 \\ y_{i0}^0 \\ \omega_i^0 \end{bmatrix} + \begin{bmatrix} \delta\omega_i \\ \delta\phi_i \\ \delta\kappa_i \\ \delta X_{i0} \\ \delta Y_{i0} \\ \delta Z_{i0} \\ \delta x_{i0} \\ \delta y_{i0} \\ \delta c_i \end{bmatrix} \tag{A.8}$$

$$\mathbf{C}_i = \tilde{\mathbf{C}}_i + \dot{\delta} \tag{A.9}$$

$$\begin{bmatrix} X_j \\ Y_j \\ Z_j \end{bmatrix} = \begin{bmatrix} X_j^0 \\ Y_j^0 \\ Z_j^0 \end{bmatrix} + \begin{bmatrix} \delta X_j \\ \delta Y_j \\ \delta Z_j \end{bmatrix} \tag{A.10}$$

$$\mathbf{P}_j = \tilde{\mathbf{P}}_j + \ddot{\delta} \tag{A.11}$$

where ω_i^0, ϕ_i^0, κ_i^0, X_{i0}^0, Y_{i0}^0, Z_{i0}^0, x_{i0}^0, y_{i0}^0 and c_{i0}^0 are initial approximations to the (unknown) actual values of the parameters ω_i, ϕ_i, κ_i, X_{i0}, Y_{i0}, Z_{i0}, x_{i0}, y_{i0} and c_{i0} for camera $\mathbf{C}_i$ and $\delta\omega_i$, $\delta\phi_i$, $\delta\kappa_i$, δX_{i0}, δY_{i0}, δZ_{i0}, δx_{i0}, δy_{i0} and δc_{i0} are the respective corrections to the approximations that are to be calculated by bundle adjustment. Similarly, X_j^0, Y_j^0 and Z_j^0 are approximations to the (unknown) actual values of the co-ordinates, X_j, Y_j and Z_j, of world point $\mathbf{P}_j$ and δX_j, δY_j and δZ_j are the respective corrections to the approximations that are to be calculated by bundle adjustment (note that the reason for keeping the camera parameters and world point co-ordinates separate will become apparent).

Newton's first order approximation, or Taylor series expansion neglecting second and higher order terms, can be used to linearise equations (A.4) and (A.5) in the vicinity of

the initial approximations, $\tilde{\mathbf{C}}_i$ and $\tilde{\mathbf{P}}_j$, and the measured image co-ordinates, $\hat{\mathbf{p}}_{ij}$, to give:

$$f_{x_{ij}}\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)+\sum_{\alpha=\omega_i,\phi_i,\kappa_i,X_{i0},Y_{i0},Z_{i0},x_{i0},y_{i0},c_i,X_j,Y_j,Z_j}\left(\frac{\partial f_{x_{ij}}}{\partial\alpha}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)\cdot\delta\alpha+v_{x_{ij}}=0 \tag{A.12}$$

$$f_{y_{ij}}\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)+\sum_{\alpha=\omega_i,\phi_i,\kappa_i,X_{i0},Y_{i0},Z_{i0},x_{i0},y_{i0},c_i,X_j,Y_j,Z_j}\left(\frac{\partial f_{y_{ij}}}{\partial\alpha}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)\cdot\delta\alpha+v_{y_{ij}}=0 \tag{A.13}$$

where $f_{x_{ij}}\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)$ and $f_{y_{ij}}\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)$ are, respectively, $f_{x_{ij}}$ and $f_{y_{ij}}$ given in equations (A.4) and (A.5), evaluated for the approximate camera parameters, $\tilde{\mathbf{C}}_i$, the approximate world co-ordinates, $\tilde{\mathbf{P}}_j$, and the measured image co-ordinates, $\hat{\mathbf{p}}_{ij}$. The partial derivatives $\left(\frac{\partial f_{x_{ij}}}{\partial\alpha}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)$ and $\left(\frac{\partial f_{y_{ij}}}{\partial\alpha}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)$ are the partial derivatives of $f_{x_{ij}}$ and $f_{y_{ij}}$ respectively with respect to the parameter α, also evaluated for the approximate camera parameters, $\tilde{\mathbf{C}}_i$, the approximate world co-ordinates, $\tilde{\mathbf{P}}_j$, and the measured image co-ordinates, $\hat{\mathbf{p}}_{ij}$.

Equations (A.12) and (A.13) can be expressed more conveniently in matrix form as:

$$\mathbf{v}_{ij}+\dot{\mathbf{B}}_{ij}\dot{\delta}_i+\ddot{\mathbf{B}}_{ij}\ddot{\delta}_j=\mathbf{f}_{ij} \tag{A.14}$$

where:

$$\underset{(2\times1)}{\mathbf{v}_{ij}}=\begin{bmatrix}v_{x_{ij}} & v_{y_{ij}}\end{bmatrix}^{\mathrm{T}} \tag{A.15}$$

$$\underset{(2\times9)}{\dot{\mathbf{B}}_{ij}}=\begin{bmatrix}\left(\frac{\partial f_{x_{ij}}}{\partial\omega_i}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \left(\frac{\partial f_{x_{ij}}}{\partial\phi_i}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \cdots & \left(\frac{\partial f_{x_{ij}}}{\partial c_i}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)\\ \left(\frac{\partial f_{y_{ij}}}{\partial\omega_i}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \left(\frac{\partial f_{y_{ij}}}{\partial\phi_i}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \cdots & \left(\frac{\partial f_{y_{ij}}}{\partial c_i}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right)\end{bmatrix} \tag{A.16}$$

$$\underset{(9\times1)}{\dot{\delta}_i}=\begin{bmatrix}\delta\omega_i & \delta\phi_i & \delta\kappa_i & \delta X_{i0} & \delta Y_{i0} & \delta Z_{i0} & \delta x_{i0} & \delta y_{i0} & \delta c_i\end{bmatrix}^{\mathrm{T}} \tag{A.17}$$

$$\underset{(2\times3)}{\ddot{\mathbf{B}}_{ij}} = \begin{bmatrix} \left(\frac{\partial f_{x_{ij}}}{\partial X_j}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \left(\frac{\partial f_{x_{ij}}}{\partial Y_j}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \left(\frac{\partial f_{x_{ij}}}{\partial Z_j}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) \\ \left(\frac{\partial f_{y_{ij}}}{\partial X_j}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \left(\frac{\partial f_{y_{ij}}}{\partial Y_j}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & \left(\frac{\partial f_{y_{ij}}}{\partial Z_j}\right)\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) \end{bmatrix} \tag{A.18}$$

$$\underset{(3\times1)}{\ddot{\delta}_j} = \begin{bmatrix} \delta X_j & \delta Y_j & \delta Z_j \end{bmatrix}^{\mathrm{T}} \tag{A.19}$$

$$\underset{(2\times1)}{\mathbf{f}_{ij}} = \begin{bmatrix} -f_{x_{ij}}\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) & -f_{y_{ij}}\left(\tilde{\mathbf{C}}_i,\tilde{\mathbf{P}}_j,\hat{\mathbf{p}}_{ij}\right) \end{bmatrix}^{\mathrm{T}} \tag{A.20}$$

Note that the superscripts "one dot" and "two dots" are used to denote the corrections to the camera parameters and world point co-ordinates respectively.

Expanding equation (A.14) for all m cameras, $\mathbf{C}_i, i = 1,2,\cdots,m$, yields:

$$\begin{bmatrix} \mathbf{v}_{1j} \\ \mathbf{v}_{2j} \\ \vdots \\ \mathbf{v}_{mj} \end{bmatrix} + \begin{bmatrix} \dot{\mathbf{B}}_{1j} & & & \\ & \dot{\mathbf{B}}_{2j} & & \\ & & \ddots & \\ & & & \dot{\mathbf{B}}_{mj} \end{bmatrix} \begin{bmatrix} \dot{\delta}_1 \\ \dot{\delta}_2 \\ \vdots \\ \dot{\delta}_m \end{bmatrix} + \begin{bmatrix} \ddot{\mathbf{B}}_{1j} \\ \ddot{\mathbf{B}}_{2j} \\ \vdots \\ \ddot{\mathbf{B}}_{mj} \end{bmatrix} \ddot{\delta}_j = \begin{bmatrix} \mathbf{f}_{1j} \\ \mathbf{f}_{2j} \\ \vdots \\ \mathbf{f}_{mj} \end{bmatrix} \tag{A.21}$$

$$\underset{(2m\times1)}{\mathbf{v}_j} + \underset{(2m\times9m)}{\dot{\mathbf{B}}_j} \underset{(9m\times1)}{\dot{\delta}} + \underset{(2m\times3)}{\ddot{\mathbf{B}}_j} \underset{(3\times1)}{\ddot{\delta}_j} = \underset{(2m\times1)}{\mathbf{f}_j} \tag{A.22}$$

Now expanding equation (A.22) for all n world point, $\mathbf{P}_j, j = 1,2,\cdots,n$, yields:

$$\begin{bmatrix} \mathbf{v}_{11} \\ \mathbf{v}_{21} \\ \vdots \\ \mathbf{v}_{m1} \\ \mathbf{v}_{12} \\ \mathbf{v}_{22} \\ \vdots \\ \mathbf{v}_{m2} \\ \vdots \\ \mathbf{v}_{1n} \\ \mathbf{v}_{2n} \\ \vdots \\ \mathbf{v}_{mn} \end{bmatrix} + \begin{bmatrix} \dot{\mathbf{B}}_{11} & & & & & & & & & & & \\ & \dot{\mathbf{B}}_{21} & & & & & & & & & & \\ & & \ddots & & & & & & & & & \\ & & & \dot{\mathbf{B}}_{m1} & & & & & & & & \\ \dot{\mathbf{B}}_{12} & & & & & & & & & & & \\ & \dot{\mathbf{B}}_{22} & & & & & & & & & & \\ & & \ddots & & & & & & & & & \\ & & & \dot{\mathbf{B}}_{m2} & & & & & & & & \\ & & \vdots & & & & & & & & & \\ \dot{\mathbf{B}}_{1n} & & & & & & & & & & & \\ & \dot{\mathbf{B}}_{2n} & & & & & & & & & & \\ & & \ddots & & & & & & & & & \\ & & & \dot{\mathbf{B}}_{mn} & & & & & & & & \end{bmatrix} \begin{bmatrix} \dot{\delta}_1 \\ \dot{\delta}_1 \\ \vdots \\ \dot{\delta}_m \end{bmatrix} + \begin{bmatrix} \ddot{\mathbf{B}}_{11} & & & & & & & & & & & \\ \ddot{\mathbf{B}}_{21} & & & & & & & & & & & \\ \vdots & & & & & & & & & & & \\ \ddot{\mathbf{B}}_{m1} & & & & & & & & & & & \\ & \ddot{\mathbf{B}}_{12} & & & & & & & & & & \\ & \ddot{\mathbf{B}}_{22} & & & & & & & & & & \\ & \vdots & & & & & & & & & & \\ & \ddot{\mathbf{B}}_{m2} & & & & & & & & & & \\ & & \ddots & & & & & & & & & \\ & & & & & & & & & \ddot{\mathbf{B}}_{1n} & & \\ & & & & & & & & & \ddot{\mathbf{B}}_{2n} & & \\ & & & & & & & & & \vdots & & \\ & & & & & & & & & \ddot{\mathbf{B}}_{mn} & & \end{bmatrix} \begin{bmatrix} \ddot{\delta}_1 \\ \ddot{\delta}_1 \\ \vdots \\ \ddot{\delta}_n \end{bmatrix} = \begin{bmatrix} \mathbf{f}_{11} \\ \mathbf{f}_{21} \\ \vdots \\ \mathbf{f}_{m1} \\ \mathbf{f}_{21} \\ \mathbf{f}_{22} \\ \vdots \\ \mathbf{f}_{m2} \\ \vdots \\ \mathbf{f}_{1n} \\ \mathbf{f}_{2n} \\ \vdots \\ \mathbf{f}_{mn} \end{bmatrix} \tag{A.23}$$

$$\underset{(2mn\times 1)}{\mathbf{v}} + \underset{(2mn\times 9m)}{\dot{\mathbf{B}}}\,\underset{(9m\times 1)}{\dot{\delta}} + \underset{(2mn\times 3n)}{\ddot{\mathbf{B}}}\,\underset{(3n\times 1)}{\ddot{\delta}} = \underset{(2mn\times 1)}{\mathbf{f}} \tag{A.24}$$

Equation (A.24) is a system of $2mn$ equations: two equations arising from each point $\mathbf{P}_j, j = 1,2,\cdots,n$ imaged by each camera $\mathbf{C}_i, i = 1,2,\cdots,m$. It is noteworthy that the vector $\dot{\delta}$, that contains the corrections to the camera parameters, is invariant to the point, *j*. In contrast, the vector $\ddot{\delta}$, that contains the corrections to the co-ordinates of the world points, is invariant to the camera *i*.

A.2.2. Additional observation equations

If measurements or observations are available for some, or all, of the world point co-ordinates, these can be incorporated into the bundle adjustment in the form of additional constraints, or observation equations. If the (unknown) actual values of the co-ordinates for world point $\mathbf{P}_j$ are X_j, Y_j and Z_j, and the measured co-ordinates of $\mathbf{P}_j$ are $\hat{X}_j$, $\hat{Y}_j$ and $\hat{Z}_j$, then:

$$\begin{bmatrix} X_j \\ Y_j \\ Z_j \end{bmatrix} = \begin{bmatrix} \hat{X}_j \\ \hat{Y}_j \\ \hat{Z}_j \end{bmatrix} + \begin{bmatrix} v_{X_j} \\ v_{Y_j} \\ v_{Z_j} \end{bmatrix} \tag{A.25}$$

$$\mathbf{P}_j = \hat{\mathbf{P}}_j + \ddot{\mathbf{v}}_j \tag{A.26}$$

where v_{X_j}, v_{Y_j} and v_{Z_j}, are the unknown residuals associated with the measurements $\hat{X}_j$, $\hat{Y}_j$ and $\hat{Z}_j$, respectively.

Combining equation (A.25) with the equation for the approximate world co-ordinates for $\mathbf{P}_j$, equation (A.10), yields:

$$\begin{bmatrix} v_{X_j} \\ v_{Y_j} \\ v_{Z_j} \end{bmatrix} - \begin{bmatrix} \delta X_j \\ \delta Y_j \\ \delta Z_j \end{bmatrix} = \begin{bmatrix} X_j^{\,0} - \hat{X}_j \\ Y_j^{\,0} - \hat{Y}_j \\ Z_j^{\,0} - \hat{Z}_j \end{bmatrix} \tag{A.27}$$

$$\ddot{\mathbf{v}}_j - \ddot{\delta}_j = \ddot{\mathbf{c}}_j = \tilde{\mathbf{P}}_j - \hat{\mathbf{P}}_j \tag{A.28}$$

Expanding out equation (A.28) for all points $\mathbf{P}_j, j = 1,2,\cdots,n$ gives:

$$\begin{bmatrix} \ddot{\mathbf{v}}_1 \\ \ddot{\mathbf{v}}_2 \\ \vdots \\ \ddot{\mathbf{v}}_n \end{bmatrix} - \begin{bmatrix} \ddot{\delta}_1 \\ \ddot{\delta}_2 \\ \vdots \\ \ddot{\delta}_n \end{bmatrix} = \begin{bmatrix} \ddot{\mathbf{c}}_1 \\ \ddot{\mathbf{c}}_2 \\ \vdots \\ \ddot{\mathbf{c}}_n \end{bmatrix} \tag{A.29}$$

$$\underset{(3n\times 1)}{\ddot{\mathbf{v}}} - \underset{(3n\times 1)}{\ddot{\delta}} = \underset{(3n\times 1)}{\ddot{\mathbf{c}}} \tag{A.30}$$

Similarly, if measurements, $\hat{\omega}_i$, $\hat{\phi}_i$, $\hat{\kappa}_i$, $\hat{X}_{i0}$, $\hat{Y}_{i0}$, $\hat{Z}_{i0}$, $\hat{x}_{i0}$, $\hat{y}_{i0}$ and $\hat{c}_{i0}$, of the camera parameters for camera $\mathbf{C}_i$, ω_i, ϕ_i, κ_i, X_{i0}, Y_{i0}, Z_{i0}, x_{i0}, y_{i0} and c_{i0} respectively, are available, and v_{ω_i}, v_{ϕ_i}, v_{κ_i}, $v_{X_{i0}}$, $v_{Y_{i0}}$, $v_{Z_{i0}}$, $v_{x_{i0}}$, $v_{y_{i0}}$ and v_{c_i}, are the respective unknown residuals in the measurements, i.e.:

$$\begin{bmatrix} \omega_i \\ \phi_i \\ \kappa_i \\ X_{i0} \\ Y_{i0} \\ Z_{i0} \\ x_{i0} \\ y_{i0} \\ c_i \end{bmatrix} = \begin{bmatrix} \hat{\omega}_i \\ \hat{\phi}_i \\ \hat{\kappa}_i \\ \hat{X}_{i0} \\ \hat{Y}_{i0} \\ \hat{Z}_{i0} \\ \hat{x}_{i0} \\ \hat{y}_{i0} \\ \hat{c}_i \end{bmatrix} + \begin{bmatrix} v_{\omega_i} \\ v_{\phi_i} \\ v_{\kappa_i} \\ v_{X_{i0}} \\ v_{Y_{i0}} \\ v_{Z_{i0}} \\ v_{x_{i0}} \\ v_{y_{i0}} \\ v_{c_i} \end{bmatrix} \tag{A.31}$$

$$\mathbf{C}_i = \hat{\mathbf{C}}_i + \dot{\mathbf{v}}_i \tag{A.32}$$

then, the following observation equation can be written for the camera parameters:

$$\begin{bmatrix} v_{\omega_i} \\ v_{\phi_i} \\ v_{\kappa_i} \\ v_{X_{i0}} \\ v_{Y_{i0}} \\ v_{Z_{i0}} \\ v_{x_{i0}} \\ v_{y_{i0}} \\ v_{c_i} \end{bmatrix} - \begin{bmatrix} \delta\omega_i \\ \delta\phi_i \\ \delta\kappa_i \\ \delta X_{i0} \\ \delta Y_{i0} \\ \delta Z_{i0} \\ \delta x_{i0} \\ \delta y_{i0} \\ \delta c_i \end{bmatrix} = \begin{bmatrix} \omega_i^0 - \hat{\omega}_i \\ \phi_i^0 - \hat{\phi}_i \\ \kappa_i^0 - \hat{\kappa}_i \\ X_{i0}^0 - \hat{X}_{i0} \\ Y_{i0}^0 - \hat{Y}_{i0} \\ Z_{i0}^0 - \hat{Z}_{i0} \\ x_{i0}^0 - \hat{x}_{i0} \\ y_{i0}^0 - \hat{y}_{i0} \\ c_{i0}^0 - \hat{c}_i \end{bmatrix} \tag{A.33}$$

$$\dot{\mathbf{v}}_i - \dot{\delta}_i = \dot{\mathbf{c}}_i = \tilde{\mathbf{C}}_i - \hat{\mathbf{C}}_i \tag{A.34}$$

Expanding out equation (A.34) for all cameras $\mathbf{C}_i, i = 1,2,\cdots,m$ gives:

$$\begin{bmatrix} \dot{\mathbf{v}}_1 \\ \dot{\mathbf{v}}_2 \\ \vdots \\ \dot{\mathbf{v}}_m \end{bmatrix} - \begin{bmatrix} \dot{\delta}_1 \\ \dot{\delta}_2 \\ \vdots \\ \dot{\delta}_m \end{bmatrix} = \begin{bmatrix} \dot{\mathbf{c}}_1 \\ \dot{\mathbf{c}}_2 \\ \vdots \\ \dot{\mathbf{c}}_m \end{bmatrix} \tag{A.35}$$

$$\underset{(9m\times 1)}{\dot{\mathbf{v}}} - \underset{(9m\times 1)}{\dot{\delta}} = \underset{(9m\times 1)}{\dot{\mathbf{c}}} \tag{A.36}$$

A.2.3. Full formulation

The full formulation of the mathematical model for bundle adjustment is obtained by stacking equations (A.24), (A.30) and (A.36) to give:

$$\begin{aligned} \mathbf{v} + \dot{\mathbf{B}}\dot{\delta} + \ddot{\mathbf{B}}\ddot{\delta} &= \mathbf{f} \\ \dot{\mathbf{v}} - \dot{\delta} &= \dot{\mathbf{c}} \\ \ddot{\mathbf{v}} - \ddot{\delta} &= \ddot{\mathbf{c}} \end{aligned} \tag{A.37}$$

$$\begin{bmatrix} \mathbf{v} \\ \dot{\mathbf{v}} \\ \ddot{\mathbf{v}} \end{bmatrix} + \begin{bmatrix} \dot{\mathbf{B}} & \ddot{\mathbf{B}} \\ -\mathbf{I} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I} \end{bmatrix} \begin{bmatrix} \dot{\delta} \\ \ddot{\delta} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \dot{\mathbf{c}} \\ \ddot{\mathbf{c}} \end{bmatrix} \tag{A.38}$$

$$\underset{(2mn+9m+3n\times 1)}{\overline{\mathbf{v}}} + \underset{(2mn+9m+3n\times 9m+3n)}{\overline{\mathbf{B}}} \underset{(9m+3n\times 1)}{\delta} = \underset{(2mn+9m+3n\times 1)}{\overline{\mathbf{c}}} \tag{A.39}$$

The goal of bundle adjustment is to solve equation (A.39) for δ, the vector of corrections to the current approximate camera parameters, ω_i^0, ϕ_i^0, κ_i^0, X_{i0}^0, Y_{i0}^0, Z_{i0}^0, x_{i0}^0, y_{i0}^0 and c_{i0}^0, and current approximate world points, X_j^0, Y_j^0 and Z_j^0. These corrections are then applied to the approximations, and equation (A.39) is reformed. The reformed equation (A.39) is solved for the new vector of corrections, δ, and so the process is repeated until the solution converges. The method of solving a general problem of the form of equation (A.40) is explored in section A.4.

A.2.4. General least squares adjustment

Let Y be a random variable, the expected value of which may be expressed as a linear function of n independent variables, $X_1, X_2, \cdots, X_n$:

$$E(Y) = a_1 X_1 + a_2 X_2 + \cdots + a_n X_n \tag{A.41}$$

If m independent measurements of Y are obtained for m different combinations of the parameters, $X_1, X_2, \cdots, X_n$, then m observation equations can be written:

$$\begin{aligned} v_1 &= a_1 x_{11} + a_2 x_{12} + \cdots + a_n x_{1n} - y_1 \\ v_2 &= a_1 x_{21} + a_2 x_{22} + \cdots + a_n x_{2n} - y_2 \\ &\vdots \\ v_m &= a_1 x_{m1} + a_2 x_{m2} + \cdots + a_n x_{mn} - y_m \end{aligned} \tag{A.42}$$

where y_i is the i-th observation of Y given that $X_1 = x_{i1}, X_2 = x_{i2}, \cdots, X_n = x_{in}$ and v_i is the residual in the observation. Equation (A.42) can be rewritten in matrix form as:

$$\mathbf{v} = \mathbf{B}\mathbf{a} - \mathbf{c} \tag{A.43}$$

where:

$$\underset{(m\times 1)}{\mathbf{v}} = \begin{bmatrix} v_1 & v_2 & \cdots & v_m \end{bmatrix}^{\mathrm{T}} \tag{A.44}$$

$$\underset{(m\times n)}{\mathbf{B}} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \cdots & x_{2n} \\ \vdots & \vdots & & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{A.45}$$

$$\underset{(n\times 1)}{\mathbf{a}} = \begin{bmatrix} a_1 & a_2 & \cdots & a_n \end{bmatrix}^{\mathrm{T}} \tag{A.46}$$

$$\underset{(m\times 1)}{\mathbf{c}} = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix}^{\mathrm{T}} \tag{A.47}$$

If there are more observations than unknown parameters ($m > n$), then a least squares adjustment can be used to find the most probable values of the residuals, $v_1, v_2, \cdots, v_m$, and the coefficients, $a_1, a_2, \cdots, a_n$. If each residual, v_i, has a normal distribution with a mean of zero and a standard deviation of σ_i, and if the observations, $y_1, y_2, \cdots, y_m$, are mutually independent, then the most probable values for $v_1, v_2, \cdots, v_m$ are those that minimise the term:

$$Q = \sum_{i=1}^{m} \left(\frac{v_i}{\sigma_i} \right)^2 \tag{A.48}$$

with respect to the unknown parameters, $a_1, a_2, \cdots, a_n$. Equation (A.48) can be rewritten as the matrix equation:

$$Q = \mathbf{v}^{\mathrm{T}}\mathbf{W}\mathbf{v} \tag{A.49}$$

where $\mathbf{W}$ is the *weight matrix* and is defined by:

$$\underset{(m\times m)}{\mathbf{W}} = \sigma_0 \begin{bmatrix} \sigma_1^{\,2} & & & \\ & \sigma_2^{\,2} & & \\ & & \ddots & \\ & & & \sigma_m^{\,2} \end{bmatrix}^{-1} \tag{A.50}$$

Note that the off diagonal elements of **W** are equal to zero because the observations, $y_1, y_2, \cdots, y_m$, are assumed to be uncorrelated.

Substituting equation (A.43) for **v** in equation (A.48) yields:

$$Q = (\mathbf{Ba} - \mathbf{c})^{\mathrm{T}} \mathbf{W}(\mathbf{Ba} - \mathbf{c}) \tag{A.51}$$

$$Q = (\mathbf{a}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}} - \mathbf{c}^{\mathrm{T}})(\mathbf{WBa} - \mathbf{Wc}) \tag{A.52}$$

$$Q = \mathbf{a}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{WBa} - \mathbf{a}^{\mathrm{T}}\mathbf{B}^{\mathrm{T}}\mathbf{Wc} - \mathbf{c}^{\mathrm{T}}\mathbf{WBa} + \mathbf{c}^{\mathrm{T}}\mathbf{Wc} \tag{A.53}$$

Since the problem is to find the vector **a** that minimises Q, then the following condition must be satisfied by the solution:

$$\frac{\partial Q}{\partial \mathbf{a}} = 0 \tag{A.54}$$

Substituting equation (A.53) for Q in equation (A.54) yields:

$$\frac{\partial Q}{\partial \mathbf{a}} = 2\mathbf{B}^{\mathrm{T}}\mathbf{WBa} - \mathbf{B}^{\mathrm{T}}\mathbf{Wc} - (\mathbf{c}^{\mathrm{T}}\mathbf{WB})^{\mathrm{T}} = 0 \tag{A.55}$$

Since **W** is a symmetric matrix, $\mathbf{W}^{\mathrm{T}} = \mathbf{W}$. Therefore, equation (A.55) can be rearranged to give:

$$\mathbf{B}^{\mathrm{T}}\mathbf{WBa} = \mathbf{B}^{\mathrm{T}}\mathbf{Wc} \tag{A.56}$$

So that:

$$\mathbf{Na} = \mathbf{k} \tag{A.57}$$

where:

$$\underset{(n\times n)}{\mathbf{N}} = \mathbf{B}^{\mathrm{T}}\mathbf{WB} \tag{A.58}$$

$$\underset{(n\times 1)}{\mathbf{k}} = \mathbf{B}^{\mathrm{T}}\mathbf{Wc} \tag{A.59}$$

A.2.5. Formation of normal equations

Applying the general equations for least squares minimisation, (A.57), (A.58) and (A.59) to the specific instance of the bundle adjustment equation, (A.39), gives:

$$\overline{\mathbf{N}}\delta = \overline{\mathbf{k}} \tag{A.60}$$

$$\underset{(9m+3n\times 9m+3n)}{\overline{\mathbf{N}}} = \overline{\mathbf{B}}^{\mathrm{T}}\overline{\mathbf{W}}\overline{\mathbf{B}} \tag{A.61}$$

$$\underset{(9m+3n\times 1)}{\overline{\mathbf{k}}} = \overline{\mathbf{B}}^{\mathrm{T}}\overline{\mathbf{W}}\overline{\mathbf{c}} \tag{A.62}$$

Equation (A.60) is known as the normal equation. The structure of the weight matrix, $\overline{\mathbf{W}}$, is described in section A.6. The full form of the $\overline{\mathbf{N}}$ matrix and the $\overline{\mathbf{k}}$ vector and the solution to (A.60) is described in section A.0.

A.2.6. Structure of the weight matrix

The structure of the vector of observation residuals, $\overline{\mathbf{v}}$, in equation (A.39) is:

$$\overline{\mathbf{v}} = \begin{bmatrix} \mathbf{v} \\ \dot{\mathbf{v}} \\ \ddot{\mathbf{v}} \end{bmatrix} \tag{A.63}$$

where $\mathbf{v}$ is the vector of residuals associated with the observation of the image co-ordinates, $\dot{\mathbf{v}}$ is the vector of residuals associated with the observation of the camera parameters and $\ddot{\mathbf{v}}$ is the vector of residuals associated with the observation of the world points. The weight matrix, $\overline{\mathbf{W}}$, corresponding to these observations is therefore a square matrix which has a block diagonal structure and consists of sub-matrices stacked in a similar manner to $\overline{\mathbf{v}}$:

$$\overline{\mathbf{W}} = \begin{bmatrix} \mathbf{W} & & \\ & \dot{\mathbf{W}} & \\ & & \ddot{\mathbf{W}} \end{bmatrix} \tag{A.64}$$

where, $\mathbf{W}$ is the weight matrix associated with the observations of the image co-ordinates, $\dot{\mathbf{W}}$ is the weight matrix associated with the observations of the camera parameters and $\ddot{\mathbf{W}}$ is the weight matrix associated with the observations of the world points.

The structure of $\mathbf{W}$ is derived as follows. If $\sigma^2_{x_{ij}}$ and $\sigma^2_{y_{ij}}$ are the variances of the measurements of x_{ij} and y_{ij} the co-ordinates of image point $\mathbf{p}_{ij}$, $\sigma_{x_{ij}y_{ij}}$ is the covariance of the measurements and σ_0^2 is a constant, then the variance-covariance matrix for the measurement is:

$$\sigma_{ij} = \sigma_0^2 \begin{bmatrix} \sigma^2_{x_{ij}} & \sigma_{x_{ij}y_{ij}} \\ \sigma_{x_{ij}y_{ij}} & \sigma^2_{x_{ij}} \end{bmatrix} \tag{A.65}$$

The corresponding weight matrix is then defined as:

$$\mathbf{W}_{ij} = \frac{1}{\sigma_0^2}\begin{bmatrix} \sigma_{x_{ij}}^2 & \sigma_{x_{ij}y_{ij}} \\ \sigma_{x_{ij}y_{ij}} & \sigma_{x_{ij}}^2 \end{bmatrix}^{-1} \tag{A.66}$$

Assuming that there is no correlation among the measurements of the individual image points, the weight matrix $\mathbf{W}$ is obtained by stacking all of the weight matrices $\mathbf{W}_{ij}$ for all n world point, $\mathbf{P}_j, j = 1,2,\cdots,n$ and all m cameras, $\mathbf{C}_i, i = 1,2,\cdots,m$ to give:

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_{11} & & & & & & & & \\ & \mathbf{W}_{21} & & & & & & & \\ & & \ddots & & & & & & \\ & & & \mathbf{W}_{m1} & & & & & \\ & & & & \ddots & & & & \\ & & & & & \mathbf{W}_{ij} & & & \\ & & & & & & \ddots & & \\ & & & & & & & \mathbf{W}_{1n} & \\ & & & & & & & & \mathbf{W}_{2n} \\ & & & & & & & & & \ddots \\ & & & & & & & & & & \mathbf{W}_{mn} \end{bmatrix} \tag{A.67}$$

Similarly, the weight matrix for the observations of the parameters for camera $\mathbf{C}_i$ is:

$$\dot{\mathbf{W}}_i = \frac{1}{\sigma_0^2}\begin{bmatrix} \sigma_{\omega_i}^2 & \sigma_{\omega_i\kappa_i} & \sigma_{\omega_i\phi_i} & \sigma_{\omega_i X_{i0}} & \sigma_{\omega_i Y_{i0}} & \sigma_{\omega_i Z_{i0}} & \sigma_{\omega_i x_{i0}} & \sigma_{\omega_i y_{i0}} & \sigma_{\omega_i c_i} \\ \sigma_{\kappa_i\omega_i} & \sigma_{\kappa_i}^2 & \sigma_{\kappa_i\phi_i} & \sigma_{\kappa_i X_{i0}} & \sigma_{\kappa_i Y_{i0}} & \sigma_{\kappa_i Z_{i0}} & \sigma_{\kappa_i x_{i0}} & \sigma_{\kappa_i y_{i0}} & \sigma_{\kappa_i c_i} \\ \sigma_{\phi_i\omega_i} & \sigma_{\phi_i\kappa_i} & \sigma_{\phi_i}^2 & \sigma_{\phi_i X_{i0}} & \sigma_{\phi_i Y_{i0}} & \sigma_{\phi_i Z_{i0}} & \sigma_{\phi_i x_{i0}} & \sigma_{\phi_i y_{i0}} & \sigma_{\phi_i c_i} \\ \sigma_{X_{i0}\omega_i} & \sigma_{X_{i0}\kappa_i} & \sigma_{X_{i0}\phi_i} & \sigma_{X_{i0}}^2 & \sigma_{X_{i0}Y_{i0}} & \sigma_{X_{i0}Z_{i0}} & \sigma_{X_{i0}x_{i0}} & \sigma_{X_{i0}y_{i0}} & \sigma_{X_{i0}c_i} \\ \sigma_{Y_{i0}\omega_i} & \sigma_{Y_{i0}\kappa_i} & \sigma_{Y_{i0}\phi_i} & \sigma_{Y_{i0}X_{i0}} & \sigma_{Y_{i0}}^2 & \sigma_{Y_{i0}Z_{i0}} & \sigma_{Y_{i0}x_{i0}} & \sigma_{Y_{i0}y_{i0}} & \sigma_{Y_{i0}c_i} \\ \sigma_{Z_{i0}\omega_i} & \sigma_{Z_{i0}\kappa_i} & \sigma_{Z_{i0}\phi_i} & \sigma_{Z_{i0}X_{i0}} & \sigma_{Z_{i0}Y_{i0}} & \sigma_{Z_{i0}}^2 & \sigma_{Z_{i0}x_{i0}} & \sigma_{Z_{i0}y_{i0}} & \sigma_{Z_{i0}c_i} \\ \sigma_{x_{i0}\omega_i} & \sigma_{x_{i0}\kappa_i} & \sigma_{x_{i0}\phi_i} & \sigma_{x_{i0}X_{i0}} & \sigma_{x_{i0}Y_{i0}} & \sigma_{x_{i0}Z_{i0}} & \sigma_{x_{i0}}^2 & \sigma_{x_{i0}y_{i0}} & \sigma_{x_{i0}c_i} \\ \sigma_{y_{i0}\omega_i} & \sigma_{y_{i0}\kappa_i} & \sigma_{y_{i0}\phi_i} & \sigma_{y_{i0}X_{i0}} & \sigma_{y_{i0}Y_{i0}} & \sigma_{y_{i0}Z_{i0}} & \sigma_{y_{i0}x_{i0}} & \sigma_{y_{i0}}^2 & \sigma_{y_{i0}c_i} \\ \sigma_{c_i\omega_i} & \sigma_{c_i\kappa_i} & \sigma_{c_i\phi_i} & \sigma_{c_i X_{i0}} & \sigma_{c_i Y_{i0}} & \sigma_{c_i Z_{i0}} & \sigma_{c_i x_{i0}} & \sigma_{c_i y_{i0}} & \sigma_{c_i}^2 \end{bmatrix}^{-1} \tag{A.68}$$

Assuming no correlation among the parameters of different cameras, the complete set of weight matrices for all m cameras, $\mathbf{C}_i, i = 1,2,\cdots,m$ is:

$$\dot{\mathbf{W}} = \begin{bmatrix} \dot{\mathbf{W}}_1 & & & \\ & \dot{\mathbf{W}}_2 & & \\ & & \ddots & \\ & & & \dot{\mathbf{W}}_m \end{bmatrix} \tag{A.69}$$

The weight matrix for the observation of the co-ordinates of world point $\mathbf{P}_j$ is:

$$\ddot{\mathbf{W}}_j = \frac{1}{\sigma_0^2} \begin{bmatrix} \sigma_{X_j}^2 & \sigma_{X_jY_j} & \sigma_{X_jZ_j} \\ \sigma_{Y_jX_j} & \sigma_{Y_j}^2 & \sigma_{Y_jZ_j} \\ \sigma_{Z_jX_j} & \sigma_{Z_jY_j} & \sigma_{Z_j}^2 \end{bmatrix}^{-1} \tag{A.70}$$

Again, assuming no correlation between the measurement of the co-ordinates of different world points, the complete set of weight matrices for all n world point, $\mathbf{P}_j, j = 1,2,\cdots,n$ is:

$$\ddot{\mathbf{W}} = \begin{bmatrix} \ddot{\mathbf{W}}_1 & & & \\ & \ddot{\mathbf{W}}_2 & & \\ & & \ddots & \\ & & & \ddot{\mathbf{W}}_n \end{bmatrix} \tag{A.71}$$

A.2.7. Solution of the normal equations

Having defined the structure of the weight matrix, $\overline{\mathbf{W}}$, equation (A.60) can then be solved for the vector of corrections, δ. Most software maths libraries now contain routines to solve linear systems of equations such as equation (A.60). However the special structure of the matrices in equation (A.60) means that it can be reduced to a form that is much more simple to solve. Firstly, it can be shown by expanding equation (A.60) right out that the equation can be partitioned into image invariant and point invariant components:

$$\begin{bmatrix} \dot{\mathbf{N}} & \mathbf{N} \\ \mathbf{N}^{\mathrm{T}} & \ddot{\mathbf{N}} \end{bmatrix} \begin{bmatrix} \dot{\delta} \\ \ddot{\delta} \end{bmatrix} = \begin{bmatrix} \dot{\mathbf{k}} \\ \ddot{\mathbf{k}} \end{bmatrix} \tag{A.72}$$

where:

$$\underset{(9m \times 9m)}{\dot{\mathbf{N}}} = \begin{bmatrix} \dot{\mathbf{N}}_1 + \dot{\mathbf{W}}_1 & & & \\ & \dot{\mathbf{N}}_2 + \dot{\mathbf{W}}_2 & & \\ & & \ddots & \\ & & & \dot{\mathbf{N}}_m + \dot{\mathbf{W}}_m \end{bmatrix} \tag{A.73}$$

$$\underset{(9\times9)}{\dot{\mathbf{N}}_i} = \sum_{j=1}^{n} \dot{\mathbf{B}}_{ij}^{\mathrm{T}} \mathbf{W}_{ij} \dot{\mathbf{B}}_{ij} \tag{A.74}$$

$$\underset{(3n\times3n)}{\ddot{\mathbf{N}}} = \begin{bmatrix} \ddot{\mathbf{N}}_1 + \ddot{\mathbf{W}}_1 & & & \\ & \ddot{\mathbf{N}}_2 + \ddot{\mathbf{W}}_2 & & \\ & & \ddots & \\ & & & \ddot{\mathbf{N}}_n + \ddot{\mathbf{W}}_n \end{bmatrix} \tag{A.75}$$

$$\underset{(3\times3)}{\ddot{\mathbf{N}}_j} = \sum_{i=1}^{m} \ddot{\mathbf{B}}_{ij}^{\mathrm{T}} \mathbf{W}_{ij} \ddot{\mathbf{B}}_{ij} \tag{A.76}$$

$$\underset{(9m\times3n)}{\mathbf{N}} = \begin{bmatrix} \mathbf{N}_{11} & \mathbf{N}_{12} & \cdots & \mathbf{N}_{1n} \\ \mathbf{N}_{21} & \mathbf{N}_{2} & \cdots & \mathbf{N}_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{N}_{m1} & \mathbf{N}_{m2} & \cdots & \mathbf{N}_{mn} \end{bmatrix} \tag{A.77}$$

$$\mathbf{N}_{ij} = \dot{\mathbf{B}}_{ij}^{\mathrm{T}} \mathbf{W}_{ij} \ddot{\mathbf{B}}_{ij} \tag{A.78}$$

$$\underset{(9m\times1)}{\dot{\mathbf{k}}} = \begin{bmatrix} \dot{\mathbf{k}}_1 - \dot{\mathbf{W}}_1 \dot{\mathbf{c}}_1 \\ \dot{\mathbf{k}}_2 - \dot{\mathbf{W}}_2 \dot{\mathbf{c}}_2 \\ \vdots \\ \dot{\mathbf{k}}_m - \dot{\mathbf{W}}_m \dot{\mathbf{c}}_m \end{bmatrix} \tag{A.79}$$

$$\dot{\mathbf{k}}_i = \sum_{j=1}^{n} \dot{\mathbf{B}}_{ij}^{\mathrm{T}} \mathbf{W}_{ij} \mathbf{f}_{ij} \tag{A.80}$$

$$\underset{(3n\times1)}{\ddot{\mathbf{k}}} = \begin{bmatrix} \ddot{\mathbf{k}}_1 - \ddot{\mathbf{W}}_1 \ddot{\mathbf{c}}_1 \\ \ddot{\mathbf{k}}_2 - \ddot{\mathbf{W}}_1 \ddot{\mathbf{c}}_1 \\ \vdots \\ \ddot{\mathbf{k}}_n - \ddot{\mathbf{W}}_n \ddot{\mathbf{c}}_n \end{bmatrix} \tag{A.81}$$

$$\ddot{\mathbf{k}}_j = \sum_{j=1}^{n} \ddot{\mathbf{B}}_{ij}^{\mathrm{T}} \mathbf{W}_{ij} \mathbf{f}_{ij} \tag{A.82}$$

Separating equation (A.72) into separate matrix equations gives:

$$\dot{\mathbf{N}}\dot{\delta} + \mathbf{N}\ddot{\delta} = \dot{\mathbf{k}} \tag{A.83}$$

$$\mathbf{N}^{\mathrm{T}}\dot{\delta} + \ddot{\mathbf{N}}\ddot{\delta} = \ddot{\mathbf{k}} \tag{A.84}$$

Equation (A.84) gives:

$$\ddot{\delta} = \ddot{\mathbf{N}}^{-1}\left(\ddot{\mathbf{k}} - \mathbf{N}^{\mathrm{T}}\dot{\delta}\right) \tag{A.85}$$

Substituting equation (A.85) for $\ddot{\delta}$ in equation (A.83) yields:

$$\mathbf{S}\dot{\delta} = \mathbf{e} \tag{A.86}$$

where:

$$\mathbf{S} = \dot{\mathbf{N}} - \mathbf{N}\ddot{\mathbf{N}}^{-1}\mathbf{N}^{\mathrm{T}} \tag{A.87}$$

$$\mathbf{e} = \dot{\mathbf{k}} - \mathbf{N}\ddot{\mathbf{N}}^{-1}\ddot{\mathbf{k}} \tag{A.88}$$

Equation (A.86) is known as the *reduced normal equation*. The block diagonal structure of the $\ddot{\mathbf{N}}$ matrix, makes its inverse straightforward to determine:

$$\underset{(3n\times 3n)}{\ddot{\mathbf{N}}^{-1}} = \begin{bmatrix} \left(\ddot{\mathbf{N}}_1 + \ddot{\mathbf{W}}_1\right)^{-1} & & & \\ & \left(\ddot{\mathbf{N}}_2 + \ddot{\mathbf{W}}_2\right)^{-1} & & \\ & & \ddots & \\ & & & \left(\ddot{\mathbf{N}}_n + \ddot{\mathbf{W}}_n\right)^{-1} \end{bmatrix} \tag{A.89}$$

Therefore:

$$\underset{(9m\times 9m)}{\mathbf{S}} = \begin{bmatrix} \dot{\mathbf{N}}_1 + \dot{\mathbf{W}}_1 - \mathbf{S}_{11} & -\mathbf{S}_{12} & \cdots & -\mathbf{S}_{1m} \\ -\mathbf{S}_{12}^{\mathrm{T}} & \dot{\mathbf{N}}_2 + \dot{\mathbf{W}}_2 - \mathbf{S}_{22} & \cdots & -\mathbf{S}_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ -\mathbf{S}_{1m}^{\mathrm{T}} & -\mathbf{S}_{2m}^{\mathrm{T}} & \cdots & \dot{\mathbf{N}}_m + \dot{\mathbf{W}}_m - \mathbf{S}_{mm} \end{bmatrix} \tag{A.90}$$

$$\underset{(9m\times 1)}{\mathbf{e}} = \begin{bmatrix} \dot{\mathbf{k}}_1 - \dot{\mathbf{W}}_1\dot{\mathbf{c}}_1 - \mathbf{e}_1 \\ \dot{\mathbf{k}}_2 - \dot{\mathbf{W}}_2\dot{\mathbf{c}}_2 - \mathbf{e}_2 \\ \vdots \\ \dot{\mathbf{k}}_m - \dot{\mathbf{W}}_m\dot{\mathbf{c}}_m - \mathbf{e}_m \end{bmatrix} \tag{A.91}$$

where:

$$\mathbf{S}_{ik} = \sum_{j=1}^{n} \mathbf{N}_{ij}\left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right)^{-1}\mathbf{N}_{kj}^{\mathrm{T}} \tag{A.92}$$

$$\mathbf{e}_i = \sum_{j=1}^{n} \mathbf{N}_{ij}\left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right)^{-1}\left(\ddot{\mathbf{k}}_j - \ddot{\mathbf{W}}_j\ddot{\mathbf{c}}_j\right) \tag{A.93}$$

Hence, having solved for $\dot{\delta}$, using:

$$\dot{\delta} = \mathbf{S}^{-1}\mathbf{e} \tag{A.94}$$

$\dot{\delta}$ can be substituted back into equation (A.85) to solve for $\ddot{\delta}$. It can be shown that equation (A.85) can be separated into independent equations for each world point $\mathbf{P}_j, j = 1,2,\cdots,n$:

$$\ddot{\delta}_j = \left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right)^{-1}\left(\ddot{\mathbf{k}}_j - \ddot{\mathbf{W}}_j\ddot{\mathbf{c}}_j - \sum_{i=1}^{m}\mathbf{N}_{ij}^{\mathrm{T}}\cdot\dot{\delta}_i\right) \tag{A.95}$$

It can be seen that the advantage of using the reduced normal equations is that matrix inversion is required for only n 3×3 matrices, i.e. $\left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right), j = 1,2,\cdots,n$, and one $9m$×$9m$ matrix, i.e. **S**. By contrast, the inversion of a $9m$+$3n$×$9m$+$3n$ matrix, i.e. $\overline{\mathbf{N}}$, is required to solve the normal equation, equation (A.60), directly.

A.3. Implementation

A.3.1. Calculation of partial derivatives

Prior to implementing the computer program to perform bundle adjustment, it was necessary to calculate the equations of the partial derivatives:

$$\left(\frac{\partial f_{x_{ij}}}{\partial\alpha}\right), \alpha = \omega_i, \phi_i, \kappa_i, X_{i0}, Y_{i0}, Z_{i0}, x_{i0}, y_{i0}, c_i, X_j, Y_j, Z_j$$

$$\left(\frac{\partial f_{y_{ij}}}{\partial\alpha}\right), \alpha = \omega_i, \phi_i, \kappa_i, X_{i0}, Y_{i0}, Z_{i0}, x_{i0}, y_{i0}, c_i, X_j, Y_j, Z_j$$

This was done using the Mathematica mathematics package. The resulting equations were then coded into the bundle adjustment software.

A.3.2. Data input

The first stage of the bundle adjustment software is to accept the input of the observed image point co-ordinates, $\hat{\mathbf{p}}_{ij}, i = 1,2,\cdots,m, j = 1,2,\cdots,n$, and the values of any of the camera parameters or world co-ordinates that have been observed, i.e. $\hat{\mathbf{C}}_i$ and $\hat{\mathbf{P}}_j$ respectively. Weight matrices, $\mathbf{W}_{ij}$, $\dot{\mathbf{W}}_i$ and $\ddot{\mathbf{W}}_j$, are also required for each set of observations, $\hat{\mathbf{p}}_{ij}$, $\hat{\mathbf{C}}_i$ and $\hat{\mathbf{P}}_j$ respectively.

Initial approximate values must be input for all of the camera parameters and world co-ordinates that have not been observed, i.e. $\tilde{\mathbf{C}}_i$ and $\tilde{\mathbf{P}}_j$. The initial approximates for those camera parameters and world co-ordinates that have been observed are set to the observed values.

A.3.3. Initialisation

At the start of each iteration of bundle adjustment, all of the matrices $\dot{\mathbf{B}}_{ij}$ and $\ddot{\mathbf{B}}_{ij}$ for each camera, $i = 1,2,\cdots,m$, and each world point, $j = 1,2,\cdots,n$, are formed by inserting the current values of $\hat{\mathbf{p}}_{ij}$, $\tilde{\mathbf{C}}_i$ and $\tilde{\mathbf{P}}_j$ into equations (A.16) and (A.18) respectively. The transpose matrices, $\dot{\mathbf{B}}_{ij}^{\mathrm{T}}$ and $\ddot{\mathbf{B}}_{ij}^{\mathrm{T}}$, are also calculated. Similarly all of the vectors $\mathbf{f}_{ij}$ for each camera, $i = 1,2,\cdots,m$, and each world point, $j = 1,2,\cdots,n$, are formed by inserting the current values of $\hat{\mathbf{p}}_{ij}$, $\tilde{\mathbf{C}}_i$ and $\tilde{\mathbf{P}}_j$ into equation (A.20).

For each camera, $\mathbf{C}_i$, for which there are parameter observations, the vector $\dot{\mathbf{c}}_i$ is calculated by subtracting the observed parameters, $\hat{\mathbf{C}}_i$, from their current approximations, $\tilde{\mathbf{C}}_i$, as shown in equation (A.34). Similarly, for each world point, $\mathbf{P}_j$, for which there is an observation, $\hat{\mathbf{P}}_j$, the vector $\ddot{\mathbf{c}}_j$ is calculated by subtracting the observed co-ordinates, $\hat{\mathbf{P}}_j$, from their current approximations, $\tilde{\mathbf{P}}_j$, as shown in equation (A.28).

A.3.4. Formation of normal equation

The normal equations are formed by calculating:

- $\left(\dot{\mathbf{N}}_i + \dot{\mathbf{W}}_i\right)$ by inserting $\mathbf{W}_{ij}$, $\dot{\mathbf{B}}_{ij}$, $\dot{\mathbf{B}}_{ij}^{\mathrm{T}}$ and $\dot{\mathbf{W}}_i$ into equation (A.74).
- $\left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right)^{-1}$ by inserting $\mathbf{W}_{ij}$, $\ddot{\mathbf{B}}_{ij}$, $\ddot{\mathbf{B}}_{ij}^{\mathrm{T}}$ and $\ddot{\mathbf{W}}_j$ into equation (A.76) and then inverting the resulting matrix.
- $\mathbf{N}_{ij}$ by inserting $\mathbf{W}_{ij}$, $\dot{\mathbf{B}}_{ij}^{\mathrm{T}}$ and $\ddot{\mathbf{B}}_{ij}$ into equation (A.78).
- $\mathbf{N}_{ij}^{\mathrm{T}}$ by transposing $\mathbf{N}_{ij}$.
- $\left(\dot{\mathbf{k}}_i - \dot{\mathbf{W}}_i\dot{\mathbf{c}}_i\right)$ by inserting $\mathbf{W}_{ij}$, $\dot{\mathbf{B}}_{ij}^{\mathrm{T}}$, $\mathbf{f}_{ij}$, $\dot{\mathbf{W}}_i$ and $\dot{\mathbf{c}}_i$ into equation (A.80).
- $\left(\ddot{\mathbf{k}}_j - \ddot{\mathbf{W}}_j\ddot{\mathbf{c}}_j\right)$ by inserting $\mathbf{W}_{ij}$, $\ddot{\mathbf{B}}_{ij}^{\mathrm{T}}$, $\mathbf{f}_{ij}$, $\ddot{\mathbf{W}}_j$ and $\ddot{\mathbf{c}}_j$ into equation (A.82).

A.3.5. Formation of reduced normal equation

The reduced normal equation is formed by calculating:

- $\mathbf{S}$ by inserting $\left(\dot{\mathbf{N}}_i + \dot{\mathbf{W}}_i\right)$, $\left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right)^{-1}$, $\mathbf{N}_{ij}$ and $\mathbf{N}_{ij}^{\mathrm{T}}$ into equation (A.90).
- $\mathbf{e}$ by inserting $\left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right)^{-1}$, $\mathbf{N}_{ij}$, $\left(\dot{\mathbf{k}}_i - \dot{\mathbf{W}}_i\dot{\mathbf{c}}_i\right)$ and $\left(\ddot{\mathbf{k}}_j - \ddot{\mathbf{W}}_j\ddot{\mathbf{c}}_j\right)$ into equation (A.91).

A.3.6. Solution of reduced normal equation

The reduced normal equation is solved for the full vector of corrections to the camera parameters, $\dot{\delta}$, by inserting **S** and **e** into equation (A.94). The vector of corrections for each camera's parameters, $\dot{\delta}_i$, is then obtained by suitably partitioning the $\dot{\delta}$ vector.

Having calculated $\dot{\delta}_i$ for each camera, the vector of corrections to each world point's co-ordinates, $\ddot{\delta}_j$, is then calculated by inserting $\dot{\delta}_i$, $\left(\ddot{\mathbf{N}}_j + \ddot{\mathbf{W}}_j\right)^{-1}$, $\left(\ddot{\mathbf{k}}_j - \ddot{\mathbf{W}}_j\ddot{\mathbf{c}}_j\right)$ and $\mathbf{N}_{ij}^{\mathrm{T}}$ into equation (A.95).

A.3.7. Iteration

At the end of each iteration, λ, new estimates of the camera parameters, $\tilde{\mathbf{C}}_i(\lambda+1)$, and the world point co-ordinates, $\tilde{\mathbf{P}}_j(\lambda+1)$, are calculated from the previous estimates, $\tilde{\mathbf{C}}_i(\lambda)$ and $\tilde{\mathbf{P}}_j(\lambda)$ respectively, and the calculated correction vectors, $\dot{\delta}_i$ and $\ddot{\delta}_j$:

$$\tilde{\mathbf{C}}_i(\lambda+1) = \tilde{\mathbf{C}}_i(\lambda) + \dot{\delta}, \quad i = 1,2,\cdots,m \tag{A.96}$$

$$\tilde{\mathbf{P}}_j(\lambda+1) = \tilde{\mathbf{P}}_j(\lambda+1) + \ddot{\delta}_j, \quad j = 1,2,\cdots,n \tag{A.97}$$

These new estimates can be used to initialise another iteration of bundle adjustment so that the estimates are further refined. Alternatively, if the solution has converged satisfactorily, (either after a specified number of iterations or when the magnitude of the correction vectors, $\dot{\delta}_i$ and $\ddot{\delta}_j$, are within specified limits) bundle adjustment exits, and $\tilde{\mathbf{C}}_i(\lambda+1)$ and $\tilde{\mathbf{P}}_j(\lambda+1)$ are returned as the final estimates of the camera parameters and world point co-ordinates.