

Alternative approaches to optophonic mappings

Michael Capp

Thesis submitted for the degree of Doctor of Philosophy
Leicester University

University College Northampton

June, 2000

UMI Number: U130244

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



UMI U130244

Published by ProQuest LLC 2013. Copyright in the Dissertation held by the Author.
Microform Edition © ProQuest LLC.

All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code.



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

Acknowledgements

There are many people I would like to thank, but foremost is Professor Phil Picton whose supervision during my years at Northampton have been a source of ideas, pertinent comments, and most of all an infectious enthusiasm for the subject. Thanks to Professor Jackie Campbell, my second supervisor, for statistical advice and moral support. Thanks also to Scott Turner, Tricia Goodchild, and the other members of the Technology & Design department.

Thanks also to Guenter and the rest of my friends who have been there when I needed them. Particular thanks go to those who have proof read this thesis. I would also like to thank the numerous guinea pigs from my village for agreeing to the optophone tests, and consequently the many headaches that followed when analysing results.

A special thanks to my family for their support and understanding, as well as the many hours spent proof reading.

Finally, I wish to thank the UK EPSRC for their funding under Research Grant Number GR/K85254.

Alternative approaches to optophonic mappings

Michael Capp

Abstract

This thesis presents a number of modifications to a blind aid, known as the video optophone, which enables a blind user to more readily interpret their local environment for enhanced mobility and navigation. Versions of this form of blind aid are generally both difficult to use and interpret, and are therefore inadequate for safe mobility. The reason for this severe problem lies in the complexity and excessive bandwidth of the optophonic output after the conversion from scene-to-sound.

The work herein describes a number of modifications that can be applied to the current optophonic process to make more efficient use of the limited bandwidth provided by the auditory system when converting scene images to sound. Various image processing and stereo techniques have been employed to artificially emulate the human visual system through the use of depth maps that successfully fade out the quantity of relatively unimportant image features, whilst emphasising the more significant regions such as nearby obstacles.

A series of experiments were designed to test these various modifications to the optophonic mapping by studying important factors of mobility and subject response whilst going about everyday life. The devised system, labelled DeLIA for the Detection, Location, Identification, and Avoidance (or Action) of obstacles, provided a means for gathering statistical data on users' interpretation of the optophonic output. An analysis of this data demonstrated a significant improvement when using the stereo cartooning technique, developed as part of this work, over the more conventional plain image as an input to an optophonic mapping from scene-to-sound.

Lastly, conclusions were drawn from the results, which indicated that the use of a stereo depth map as an input to a video optophone would improve its usefulness as an aid to general mobility. For the purposes of detecting and determining text or similar detail, either a plain unmodified image or some form of edge (depth) image were found to produce the best results.

Contents

Acknowledgements	i
Abstract	ii
Contents.....	iii
List of Figures	vii
List of Tables.....	x
1 Introduction.....	1
2 An overview of the blind aid	4
2.1 Origins of the electronic blind aid	4
2.2 An end to the reading aid.....	10
2.3 Blind mobility aids	12
2.3.1 Echolocation devices.....	12
2.3.2 Optophonic systems and mappings	17
2.4 The video optophone	22
2.4.1 Introduction	22
2.4.2 Modifications	25
2.5 Summary.....	26
3 A simplistic approach to information processing and reduction.....	28
3.1 Sound outputs	29
3.2 Image processing	32
3.2.1 Colour reduction.....	34
3.2.2 Noise reduction	36
3.2.2.1 Mean filter	36
3.2.2.2 Gaussian filter	37
3.2.2.3 Median filter.....	38
3.2.3 Edge & line detection.....	42
3.2.3.1 Edge detectors	43
3.2.3.1.1 Gradient operator	43
3.2.3.1.2 Sobel & Prewitt operators.....	44
3.2.3.1.3 Laplacian operator	46
3.2.3.2 Line detectors	51
3.2.4 Time varying edge detection	56
3.3 Summary.....	59

4	Stereo processing of images.....	61
4.1	Background to stereovision	63
4.2	Stereo techniques	65
4.2.1	Intensity depth maps.....	65
4.2.2	Edge depth maps	69
4.2.3	Matching constraints	75
4.3	Development of stereo techniques.....	77
4.3.1	Triclops camera system.....	77
4.3.2	Stereo depth maps	78
4.3.3	Stereo edge depth maps.....	80
4.3.4	Fast stereo edge depth maps.....	82
4.3.4.1	Algorithm	83
4.3.4.2	Limitations and possible remedies	84
4.3.4.3	Results	86
4.3.5	Stereo cartooning.....	89
4.4	Summary.....	94
5	Optophonic performance – Stereo versus monocular vision	96
5.1	Evaluation of depth maps	96
5.1.1	False Positive Fraction	97
5.1.2	Stereo algorithm	99
5.1.3	Stereo performance	101
5.2	Optophonic testing.....	106
5.2.1	Initial experiments.....	106
5.2.1.1	Test1 – Frame rate.....	107
5.2.1.2	Test2 – Obstacle avoidance.....	108
5.2.2	Further assessment	110
5.2.2.1	Test3 – Preliminary optophonic evaluation.....	110
5.2.2.1.1	Method.....	110
5.2.2.1.2	Results & analysis.....	113
5.2.2.2	Test4 – Further optophonic testing.....	120
5.3	Summary.....	127
6	Other approaches.....	141
6.1	3-D soundscapes	141
6.2	The Neurophone	142

6.3	Foveated images	143
6.4	Monocular stereovision	147
6.4.1	Translational fundamental matrix	147
6.4.2	Derivation of the translational fundamental matrix.....	153
6.4.3	The Hough solution.....	160
6.4.4	Depth from defocus.....	163
6.5	Summary.....	168
7	Conclusions.....	171
7.1	In the beginning.....	171
7.2	The approach	172
7.3	Developments	173
7.4	Experimentation & analysis.....	175
7.4.1	Initial experimentation	175
7.4.1.1	Test 1 – Frame rate.....	175
7.4.1.2	Test 2 – Obstacle avoidance.....	176
7.4.2	Further assessment	178
7.4.2.1	Test 3 – Preliminary optophonic evaluation.....	178
7.4.2.2	Test 4 – Further optophonic testing.....	179
7.5	Future development	180
7.6	Summary.....	182
	Appendix 1. Timeline of devices for the blind.....	184
	Appendix 2. Test data.....	200
A2.1	Test 3 – Preliminary test data	200
A2.2	Test 4 – Test data.....	210
	Appendix 3. Hardware & software considerations.....	251
A3.1	Hardware	251
A3.2	Software.....	252
A3.2.1	Software compilers.....	252
A3.2.2	Test programmes	253
A3.2.2.1	Stereo algorithm programme.....	254
A3.2.2.2	Optophone test programme	257
A3.2.2.3	Real-time stereo optophone programme	260
A3.3	Hardware & software problems.....	262

A3.3.1	Camera misalignment.....	262
A3.3.2	Programme development.....	265
A3.3.3	Optimisation.....	267
Appendix 4. Publications.....		269
References.....		302

List of figures

2.1	A blind man using the ‘Exploring Optophone’	5
2.2	Notes produced by Fournier d’Albe’s optophone	7
2.3	Model of type-reading Optophone	8
2.4	Miss Mary Jameson at the Reading Optophone	9
2.5	A simplified diagram of Meijer’s optophone system	23
2.6	The basic processes involved in capturing images and the respective conversions to sound for both the original optophone and the proposed stereo optophone	25
3.1	Representation of dichotic presentation	30
3.2	Angular displacement of a virtual sound source	31
3.3	(a) – Sample 256x256-pixel image. (b) – Example of the general input to the modern optophone – 64x64 pixels and 16 grey levels	34
3.4	(a) – The 4 colours of best fit for the whole image. (b) – The 4 colours of best fit for individual vertical columns.....	35
3.5	(a)-(h) – Result of applying the median, mean, and Gaussian filters to a test image of an office scene	40
3.6	A graphical representation of the 5x5 Gaussian filter with $\sigma^2 = 2$	38
3.7	3x3 neighbourhood median filter	39
3.8	The arrangement of pixels under consideration when using the 3x3 Sobel operator.....	45
3.9	Convolution masks for the Sobel edge operator.....	45
3.10	(a) – Edge in the image function $f(x, y)$. (b) – First derivative of the image function. (c) – The zero crossing of the second derivative.....	47
3.11	Application of the Laplacian operator to an image edge.....	47
3.12	(a)-(h) – Result of applying the Gradient, Sobel, and Laplacian edge operators to a test image of an office scene.....	49
3.13	(a) – Image line passing through three defined points. (b) – Classical Hough transform into $m-c$ space. (c) – Transformation of edge coordinates into $\rho - \theta$ coordinates. (d) – The modified Hough space	52
3.14	Modified Hough domain	53
3.15	(a) – Simple scene with two objects. (b) – Expected result of the Hough transform	54
3.16	(a)-(d) – Result of applying the Hough line detector to two sample images.....	56
3.17	(a)-(g) – Result of applying motion detectors to sample images.....	58
4.1	Epipolar constraint for a pair of stereo cameras	64

4.2	(a)-(i) – Result of applying intensity and edge based stereo algorithms to a sample image pair.....	72
4.3	Determining the cyclopean separation for two matching points within a stereogram	76
4.4	(a)-(c) – Result of applying Triclops stereo algorithms to a sample image set	79
4.5	Stereo matching.....	81
4.6	(a) – Two camera edge depth map. (b) – Three camera edge depth map.....	82
4.7	(a)-(e) – Result of applying the rapid stereo edge algorithm to a sample image set.....	88
4.8	(a)-(f) – Result of applying the stereo cartooning algorithm to two sample image sets	93
5.1	(a)-(z) – Set of depth maps generated with the various stereo edge algorithms tested during the research with a sample image set.....	129
5.2	(a)-(z) – Set of depth maps generated with the various stereo edge algorithms tested during the research with a second sample image set.....	133
5.3	Diagram showing the regions that cannot be matched due to lack of stereo overlap.....	100
5.4	Stereo matching without the removal of previously matched edge pixels	102
5.5	(a) & (b) – Result of applying the PMF constraint to a stereo edge algorithm with two sample image sets.....	137
5.6	Screenshot of the programme used to test a subject’s ability to perceive the location of two moveable dots by sound alone at high frame rates.....	107
5.7	(a)-(p) – Sample images used for evaluating optophonic techniques with DeLIA.....	138
5.8	(a)-(d) – Graphical representations of the performance of seven subjects during optophonic tests with plain unmodified images and the cartoon depth images.....	114
5.9	(a)-(d) – Graphical representations of the performance of twenty subjects during further optophonic tests with plain unmodified images and the cartoon depth images.....	123
5.10	A scatter graph comparing the number of required repetitions to complete the optophone test with the unmodified images and the cartoon depth images	126
6.1	A possible resolution setting for image foveation with central fixation.....	144
6.2	Simplified graphical representation of the process using foveated pyramids to create foveated images	145
6.3	(a)-(d) – Two samples of 4 layer centrally foveated images	146
6.4	(a)-(h) – The use of the Moravec interest operator and pseudo-normalised correlation to determine camera motion between frames.....	151
6.5	(a)-(i) – The use of the Moravec interest operator and pseudo-normalised correlation to determine the translation fundamental matrix enabling the generation of an edge depth map.....	157

6.6	(a)-(e) – Basic line depth map formed from the information obtained by applying the Hough line detector to a sample image set.....	162
6.7	Simplified representation of the effect that blurring has on image edges	164
6.8	(a)-(d) – An edge depth map created by the application of differing sized Laplacian edge operators to a blurred image.....	166
A3.1	Diagram of the Triclops camera system.....	251
A3.2	Screenshot of the programme written to test various stereo algorithms and other image processing techniques.....	254
A3.3	(a)-(c) – Screenshots of the stereo algorithm test programmes’ options dialog box.....	255
A3.4	(a)-(d) – Screenshots of the optophone test programme used to evaluate the performance of subjects with various optophonic modifications.....	258
A3.5	(a) & (b) – Screenshots of the real-time optophone programme.....	260
A3.6	(a) – Three-camera depth map showing a large number of errors due to camera misalignments. (b) – Three-camera depth map after the necessary corrections have been made.....	263
A3.7	(a)-(c) – A set of three Triclops images (stereo images) with marked features used to determine camera misalignments.....	263
A3.8	(a) & (b) – Diagrams representing the camera misalignment calculated from figures (A3.7a)-(A3.7c).....	265

List of tables

3.1	Values obtained from the Gaussian function, equation [3.5], with i & j in the range $[-2,2]$ and $\sigma^2 = 2$, followed by multiplying out and rounding to the nearest integer.....	37
5.1	Values obtained for the False-Positive-Fraction when used to assess the stereo edge algorithm labelled AIR1C0S0T0	99
5.2	Table of results for the False-Positive-Fraction when used to assess the performance of a stereo edge algorithm with a number of different matching constraints on an image set of a lamp-office scene	104
5.3	Table of results for the False-Positive-Fraction when used to assess the performance of a stereo edge algorithm with a number of different matching constraints on an image set of two large letters ('A' & 'B')	105
5.4	Results obtained during initial testing on several volunteer subjects using the images shown in figures (5.7a)-(5.7p).....	113
5.5	Data extracted from table (5.4), indicating the number of repetitions required by seven subjects to complete the optophone test, with values required for performing the <i>t-test</i>	118
5.6	Results obtained after applying both the <i>F-test</i> & <i>t-test</i> to the data shown in table (5.4), which was gathered during preliminary tests with the optophone programme.....	119
5.7	Results obtained during optophonic testing on twenty volunteer subjects using the images shown in figures (5.7a)-(5.7p).....	122
5.8	Results obtained after applying both the <i>F-test</i> & <i>t-test</i> to the data shown in table (5.7), which was gathered during the second set of tests with the optophone programme.....	125
6.1	Pixel intensities used to generate a depth map, figure (6.8d), through image blurring	165
A1	Timeline of devices for the blind.....	185
A3.1	Coordinates of the crosses marked in figures (A3.7a)-(A3.7c) used to calculate camera misalignment.....	264

1. Introduction

The role of a blind mobility aid is considered in an attempt to derive the factors necessary to create the ideal mobility aid for the blind individual. Topics taken into account include the ease of use, and effectiveness of the device to improve a blind user's mobility.

This is followed with the aims, and justification for this research, and a brief description on the topics and techniques employed in an attempt to complete the described tasks. Answers are given to why further research was undertaken in attempt to revamp an old idea, and what gains could be achieved in the process.

In Britain alone, it is estimated that there are in the region of one million registered blind or partially sighted people. A large percentage of these are unable to read text, other than via Braille or the Moon system [Moo86], or navigate in the same way as a sighted person. Of course the members of the blind community generally cope extremely well. They have no other choice. Hence, it is easy to see why over the last century there has been great interest in producing devices for the blind with the hope of at least partially solving the problems that the blind community face everyday.

When considering blind mobility aids, various questions spring to mind about what they actually do. Obviously, the perfect blind aid would enable a blind person to regain, if originally lost, their sight. Presently, even partial restoration of sight for the majority of the blind community is not possible, although this might one day be within our grasp. For the time being alternative solutions or compromises have to be made.

Most current mobility aids for the blind convert a visual image into either a tactile pattern or an audio output, which with training a blind user can learn to use to improve their mobility. However, these devices often fall well short of providing the blind user full mobility with the same efficiency as a sighted person. Furthermore,

most mobility aids are considered to be tiring to use due to the quantity of information they relay to the user.

This research was intended to investigate a blind mobility aid known as the optophone, to try with modern day computers and video technology to produce a better, more adaptable mobility aid. The hope being to diminish, or even, eliminate some of the more severe problems encountered when using such a device. If unsuccessful, the next step would be to attempt to determine the reasons for failure, and to propose a potential alternative solution.

As a starting point the research would study past failures and successes made through history to blind aids. This might allow potential weaknesses to be excluded from future devices, in favour of emphasising the strengths.

The present day optophonic mapping [Mei92a] was found to be a suitable working frame on which to generate a new software implementation of a blind mobility aid. Not only could this allow testing to be undertaken on a portable computer, but it would also provide an easy method for making rapid alterations to the algorithms under investigation. Once acceptable optophonic models were implemented in software, the efficiency of the algorithms and mappings could be tested on both sighted and visually impaired volunteers.

Before proceeding it should be mentioned that the original goals of the research were to investigate and implement a real-time version of the video optophone. Experimental data would then be gathered through the use of sighted and non-sighted volunteers. However, the proposed research was redirected to study methods for improving the optophonic scene-to-sound process, for later testing on volunteers. The reasons for this change were threefold.

Firstly, during the initial stages of the research whilst studying previous attempts at optophonic mappings, Peter Meijer created a real-time software optophone [Mei00]. Secondly, while carrying out background research it became apparent that the optophone had a larger history than was previously believed. It was felt that a thorough search through the more historic devices (a detailed list can be found in

Appendix 1) was worth pursuing, as this could help in finding ideas on how the optophone could be improved [CapPic00b]. Lastly, due to rapidly changing technology, time was frequently spent implementing an aspect of the optophone only to find a new piece of hardware or software had been developed that did a better job.

The study of the optophonic mapping was separated into two areas: image processing prior to conversion to sound, and the method of generating sound to provide directional information. A third stage also exists, which is the conversion from image to sound. However, the optophonic mappings used for this purpose are well accepted and have previously been investigated in some considerable depth by Adrian O'Hea [Ohe94], so were not considered in quite as much detail.

Whilst considering the image processing stage it was believed that a great deal of time could be saved if rather than testing each method on volunteers (which would obviously require a period of training and evaluation for each system) it would be inspected visually. Since for a sighted person it is easier to interpret a scene visually rather than via the sound output from the optophone (mainly due to the reduction of quality and alterations caused while converting from scene to sound), the following conclusion was made. If a sighted person were unable to accurately comprehend the modified optophonic image prior to conversion to sound, then for most practical cases a non-sighted person would have even more difficulty attempting to understand the equivalent sound output.

Once a suitable method for stressing important features within an image scene (whilst fading less critical features) was found through visual inspection then it could be applied and tested on sighted and non-sighted volunteers.

2. An overview of the blind aid

A concise history of blind aids is given, with particular emphasis on optophone like devices, which extends back more than a century to the earliest blind aids. This section includes details on the various modifications made to both the reading and mobility aids over the past 100 years, and in particular over more recent years with the rapid advancements in both computer and video technologies.

The blind aid known as the optophone, which has in the past been used as both a very basic mobility aid, and as a rather competent reading aid is described with its progress from Fournier d'Albe to Peter Meijer's present day version of the optophonic mapping. Included is a description of the workings of the modern day equivalent of the device, and its shortcomings that have, in the past, led to its general lack of use and overall obscurity from the public eye.

Conclusions are made upon the almost identical aspects of each blind aid and their respective image to sound or image to haptic mappings that have often led to their eventual downfall for large-scale public use, and how, with the modern day computer, they may to a greater extent be circumnavigated.

Aims of the research were thus re-evaluated, with the concept of stereovision being incorporated. It was believed that a stereo depth map would not only be successful in suppressing the quantity of less important features contained within an image of a real world scene, but that it would also provide the user with invaluable information upon the relative distance to possible obstacles in their immediate vicinity.

2.1. Origins of the electronic blind aid

Over the years there has been a wealth of research into aiding the blind and there have been many devices created to convert light energy into pulses of electricity. Some of these devices have been specifically aimed towards sensory substitution - in

particular, reading and navigation aids for the blind [See in particular – BenBen63, Beu47, BorUlr97, CapPic00b, CooGaiNye84, Fis76, Fou24, Kay84, Kur81, Mei92a, Mei92b, NieMahMea87, ShoBorKor98, and WarStr85]. For a historical breakdown of the more significant research into blind mobility and reading aids see Appendix 1.

Back in 1817, the element Selenium was discovered by Berzelius. In the following years, it became apparent that selenium was photosensitive, reacting to light in such a way as to vary its conductivity. Then in 1880, Graham Bell invented the Photophone, which conveyed speech along a beam of light to a selenium receiver. Later, in the same year, Perry and Ayrton proposed the concept of ‘electric vision’. A method by which they believed an array or mosaic of photocells (selenium) could be used to capture a representation of a scene. [Fou24]



Figure (2.1) – A blind man using the ‘Exploring Optophone’. Fournier d’Albe can be seen standing on the left of the picture. [Fou24]

It is important to realise that at that time there was no television. A number of scientists and technologists were working on the conversion of light into electricity. One such person was Dr E. E. Fournier d'Albe who in 1910 was appointed Assistant-Lecturer in Physics at the University of Birmingham in England. It was here that he set up a laboratory to look into the properties of selenium and where the word optophone was first used in 1912 [Fou24].

Although Fournier d'Albe did not invent the first television, he did discover that it was possible to listen to light by attaching a telephone receiver to a cell formed of Selenium. It was this discovery that led to the 'Exploring Optophone'. Unbeknown to him at the time, Fournier d'Albe was not the first to discover this.

A few devices had already been invented as aids for the blind, such as the Elektroftalm [StaKul63] and the Photophonic book [CooGaiNye84, Tur02]. The Elektroftalm, originally created in 1897 by Noiszewski, was a simple mobility aid for the blind. It used a single selenium cell that was placed on the forehead to control the intensity of a sound output, thus allowing a blind person to distinguish between light and dark. The Photophonic book on the other hand was a reading aid for the blind, which was created in 1902 by V. de Turine [Tur02]. This device required a specially prepared book that used a series of transparent squares to represent letters of the alphabet. A page from this book would be passed under a light. The light passing through the transparent squares would fall on a selenium cell, effectively modulating the electric current to a speaker.

Fournier d'Albe's first optophone (figure (2.1)), the 'Exploring Optophone', was very similar to the Elektroftalm. It converted light into sound via a selenium detector, the sound being directly proportional to the intensity of the light falling on the detector. Fournier d'Albe believed it would be far more useful if the Exploring Optophone could detect the contrasts of neighbouring objects rather than trying to gauge the brightness of surfaces as they came into range. With this in mind, he modified his device so that it contained two adjoining selenium cells, which would be simultaneously exposed to the light source (object). Through the use of a Wheatstone bridge it was then possible to detect an object's boundary or edge [Fou24].

The Exploring Optophone, intended as a mobility aid for blind people, was first shown in 1912 at the Optical Convention, held in the Science Museum at South Kensington. Although this device received a lot of good press coverage, it was criticised by Sir Washington Ranger, a blind solicitor, who said, “The blind problem is not to find lights or windows, but how to earn your living.” As a result of this criticism Fournier d’Albe turned his attention away from mobility aids, and in 1913 invented the first Reading Optophone. This device, first viewed in action by the public at the Birmingham meeting of the British Association on September 11, 1913, could be used to scan transparent letters about 8 centimetres in height. The letters, which were printed on a transparent material such as gelatine, were passed between the selenium detector and a light source, similar to the Photophonic book. [Bar21]

The device itself was attached to a Reed telephone receiver, invented by Sidney G. Brown in 1912. This receiver was superior to the Bell telephone receiver, and was capable of detecting currents of less than a millionth of an ampere, provided they were regularly interrupted. To achieve this Fournier d’Albe used a rotating cog (spinning between 20 and 30rps) containing slots that allowed the light to pass through at regular intervals. In one such device, a line of eight slots was used to generate eight dots of light with frequencies set to produce a diatonic scale (the eight different frequencies having the following ratios – 24, 27, 30, 32, 36, 40, 45, 48) [Fou24]. In later devices this was reduced to rows of 5 beams of light (figure (2.2)). Different notes being heard from the telephone receiver depending on the amount and position of the light passing through the transparent letters.

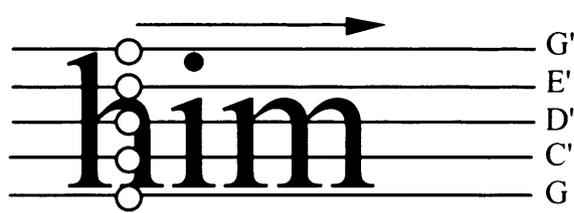


Figure (2.2) – Notes produced by Fournier d’Albe’s optophone. [EleRev21, CooGaiNye84]

After the apparent success of his first reading optophone, Fournier d’Albe set about building the ‘final stage’ of the solution to blind reading, namely to bring ordinary printed matter within the scope of the blind population [EleRev21, Fou14,

SciAme20]. Although this problem was far more difficult than first thought. For instance, inaccuracies could occur due to the limited amount of light reflected from the smaller ink-printed text.

To overcome this problem, Fournier d'Albe needed to bring the detector closer to the printed text. For this to be achieved, the detector had to be perforated to allow the light source to be shone through the selenium cell from behind, whilst the cell itself received any reflected light from the page (figure (2.3)).

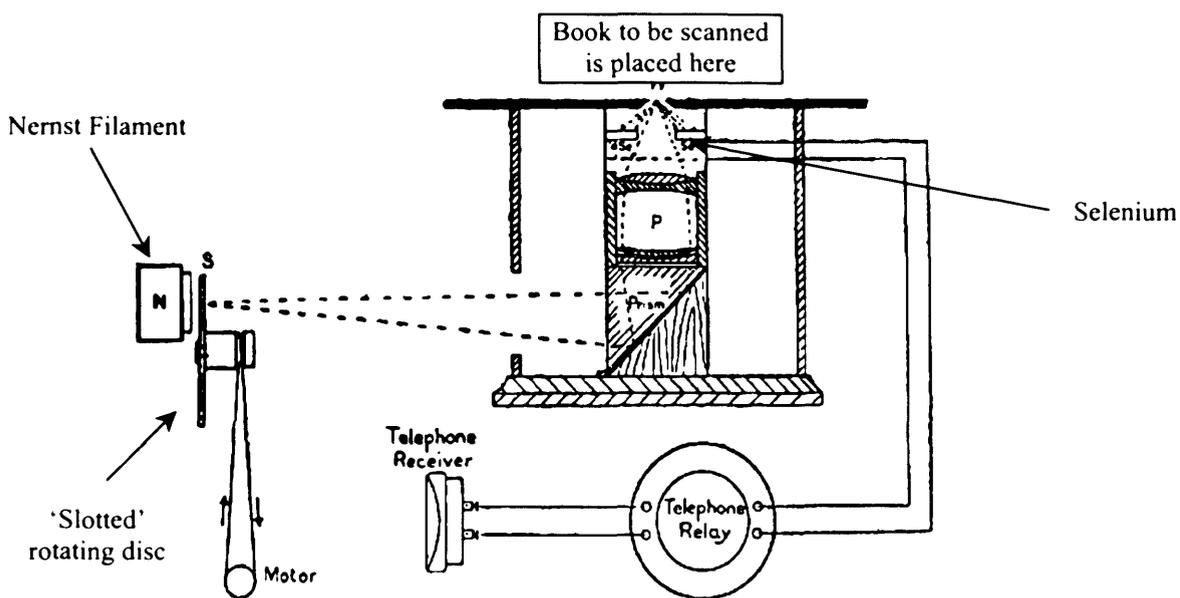


Figure (2.3) – Model of type-reading Optophone. [Bar30, Fou24]

As can be seen in figure (2.4), a book would be placed face down on top of the device upon a glass plate. The user would then be required to adjust the device until the first line of text was located. The detector would then scan each line of text, with the user's guidance, producing higher intensity sounds for the lighter areas, and no sound for the black characters. This optophone used a vertical row of five spots of light. When the light passed over the page, the reflection varied according to the character, or lack of character scanned. The reflected light was detected by a row of selenium cells, which modulated the sound output, each spot of light producing a different musical note (figure (2.2)). The idea was that a musical 'motif' would be produced

from the text, which would be far more pleasing to the ear than a set of unfamiliar sounds.

This modified optophone was first exhibited, by Fournier d'Albe, in May and June of 1914. The exhibition provided the scientific community with a demonstration of the 'type-reading optophone', showing how the device could be used to read newspaper type only five millimetres in height.



Figure (2.4) – Miss Mary Jameson at the Reading Optophone [Fou24].

There was an inevitable gap in the research during the First World War. However, in 1918 the firm of Barr & Stroud took over and developed the 'black sounding' optophone. This was accomplished by using two selenium photocells. One in constant view of the light source, the other exposed to the reflected light. The outputs from the two cells were then balanced so that the combined signals negated when subjected to a blank white page. Hence, the telephone receiver would only emit sound when the detector passed under black text. [Bar30, Beu47]

Due to the lower ratio of black print to blank paper on a page, the black sounding optophone created fewer sounds and was considered to be easier to learn. It was also found that the six scanning beams this device used could fairly accurately portray any printed character. One expert student of the device, Miss Mary Jameson (figure

(2.4)), attained a reading speed of sixty words a minute, although the average was generally far lower. At this stage the optophone seemed on the verge of success. It was manufactured in significant numbers, and had a great deal of publicity, even appearing on the cover of Scientific American in 1920 [SciAme20]. However, by 1924 it had all but disappeared, apparently due to its difficulty of use. It required much training, was stressful to use, reading speeds were slow, and its cost at this time was an expensive £100 [Beu47].

2.2. An end to the reading aid

Over the past century many researchers have attempted, often in vain, to improve or perfect the principles of the blind reading aid based upon the original optophone. However, the majority of devices have all encountered the same difficulties related to the repetitive nature of the optophonic output, and thus have generally been neglected soon after their creation. Until just after the Second World War when the Veteran's Association in America commissioned Haskins Laboratories [CooGaiNye84] to investigate the reasons for the optophone's failure as a blind reading aid. This is considered to be one of the most significant pieces of research yet to have been undertaken on any form of blind aid. During the 30 years in which Haskins Laboratories worked on the project they investigated numerous variations of the original optophone, even exploring the use of different optophonic outputs in an attempt to mimic the nature of the spoken word.

Whilst attempting to determine why the optophone had failed in its quest for widespread use they noticed a number of inherent limitations, which greatly impeded the success of a user when reading with the device:

- It was often difficult to distinguish between certain letters in the alphabet, such as 'a', 'e', 'o', and 'c'. However, it was claimed that it would become possible for the user to identify an unknown letter by listening to the 'motif' for the whole word, and with further use, to identify whole words by listening to rapidly played patterns of sound [Bar21]. This was shown to be feasible when using the

optophone at low speeds, but for greater speeds this required an enormous amount of training, concentration, and effort on the part of the listener, and in general was near impossible.

- Confusions occurred through misalignment of the text with the optophone's optical reader. Even though the original type-reading optophone had an ingenious mechanism for aligning the book to be read, it still required a fair amount of effort to correctly position the text, which had to be repeated for every page.
- A more severe problem that revealed itself was the inherent limitations of the human auditory system. When listening to a series of clicks or sounds with a low repetition rate, the human ear can readily distinguish every sound as a separate event. However, as the pulse rate is increased, the sound is perceived to transform from a buzz to a single tone of increasing pitch. This effect is first noticed when the repetition rate reaches about 20Hz. With the optophone there are on average 3 different chords per character, and approximately 5 or 6 characters per word. The maximum speed that a human could feasibly listen to an optophone and still be expected to accurately interpret the sounds would be about one word per second. When considering Morse code the researchers at Haskins laboratories realised that even this relatively slow speed would be incredibly difficult to attain. Morse code is a close parallel to the optophone's output, and has a maximum limit of 60 words a minute. Most users of Morse have difficulty in achieving speeds above 30 words a minute, which is far slower than an average sighted reader.

The researchers at Haskins Laboratories soon realised that one of the main reasons why people can accurately perceive speech at comparatively high speeds is due to the way in which the spoken word is formed. Each word is made-up of, on average, only 3 or 4 sounds (phonemes), which are formed from a combination of several letter sounds. By contrast, the optophone uses 3 or 4 sounds per character. With this new knowledge, the study concentrated on phonemes and the search for a device that could recognise and then synthesise sounds that would represent not just individual letters, but rather groups of letters (and words).

Eventually the group came to the conclusion that the ideal reading aid for the blind would be an optical character reader (OCR) linked to a speech synthesiser. By the

time that this conclusion had been made (1974) such devices were technologically possible. This was the final blow to the optophone as a reading aid for the blind. However, some research still continues, particularly in Japan where the reading of Kanji pictograms is difficult for OCR.

2.3. Blind mobility aids

Whilst the research into Blind reading aids generally ceased during the 70's, the search for an efficient Blind mobility aid continued, either in the form of the dedicated researcher, or through individuals simply toying with ideas. Some of the more obscure concepts and devices were doomed to fail from the start, while others have continued to develop to the present day. Generally the Mobility aid, or Obstacle detector, can generally be placed into two categories: - Those that work on the principles of echolocation, transmitting and then receiving reflected signals, or those that capture images directly from a video camera and then transform the scene into auditory and/or tactile forms.

2.3.1. Echolocation devices

The former category actually originated pre-1950's with echolocation devices such as the 'Clicker' [Kay84], invented by R. L. Beurle, which emitted audible 'clicks' allowing the user to locate large objects via echoes. Once a subject became proficient with the 'Clicker' they were able to use their own footfalls as a suitable sound source. Similarly, during the 1950's, the Long Cane was introduced, and could be used not only to detect objects by contact, but also by the echo generated from repeatedly tapping the cane.

Since the 1940s there have also been a number of electronic mobility aids that have been proposed. One of the more bizarre devices was the Radioactive Guider [Beu47]. This device produced a 'beam of radiations' by using a small particle of radioactive material. The radiation reflecting off a nearby object could be detected via a Geiger counter, and used as an indication of the distance to that object.

Since then, probably the best known of the transmission and detection devices are the laser cane [BenAliSch73] and the binaural/trinaural Sonicguides (so called for the number of ultrasonic receivers they employ) [IfuSasPen91, Kay84, ShoBorKor98, WarStr85 – pp. 1-12]. The Laser Cane, one version of which was introduced in the early 1970's by Benjamin *et al* (see also [BenBen63] for details of earlier work into radar type devices), emitted three beams of laser light to detect objects in the user's path [Fis76, Kay84]. On encountering an object any reflected rays would be detected and used to generate various signals to the user, to alert them of the object's presence and approximate location. One beam would be directed towards the ground to search for any ground-based objects, resulting in the generation of a low-pitched tone. The centre beam warns of any objects directly in front of the user by stimulating the index finger of the hand holding the cane. The last beam generates a high-pitched tone for objects around head height. To gain an impression of their surroundings a user would twist the cane from side-to-side.

The Sonicguides, by L. Kay, were devised from his earlier device called the Sonic Torch [Bra85, Kay84, WarStr85 – pp.1-12], which was a small torch-like device that used an ultrasonic beam to identify the distance to any object in its line-of-sight in the same way as a bat. It was intended that the blind would use the device in the same way as the sighted would use a torch at night to reveal a small area of the scenery. The Sonic Torch used an audible output, with the pitch and timbre corresponding to the range and variation in target surface texture, respectively. Unfortunately, this device had two drawbacks that caused Kay to rethink his strategy. Firstly, the torch could not be used simultaneously with the more popular long cane, since the Sonic Torch often perceived the movement of the long cane in front of the user as an obstacle. Secondly, although the sonic torch could give some impression of object texture, it was not designed to search for landmarks.

Kay's Sonicguide overcame the above limitations with a wider field of view, using echolocation to present the user with a binaural audio representation of their surroundings. The device, which was built around a pair of glasses and is worn as such, has in the case of the binaural aid, a central ultrasonic transmitter of cone 55-60°, and an ultrasonic receiver on either side of the transmitter [Kay84,

ShoBorKor98]. The trinaural system uses the same principles as the binaural device, with one main exception. It uses a third, central, narrow-field receiver in a similar fashion to that of the fovea in vision, which greatly enhances spatial resolution in the area of focal attention. Trials on young blind children, from the ages of one upwards, have shown promise. Using the device they can learn to partially replace their lost sense of sight, to the extent that the subjects demonstrated an improved spatial understanding and awareness. [HilDodHilFox95]

Recently, a group of researchers have produced the GuideCane [BorUlr97] and NavBelt [BorUlr97, ShoBorKor98]. Both of these systems work on ultrasonics, however the two devices are used in completely different ways.

The GuideCane can almost be thought of as a robotic guide dog. Held out in front of the traveller in a similar fashion to that of the white (long) cane, it has a series of ultrasonic sensors housed in a unit at the end of the cane, which is supported by two guide wheels. This unit also contains a steering servomotor that is operated by an onboard computer. When an obstacle is detected in the traveller's path, the computer calculates a new 'safe' course and alters its steering servomotor. Assuming this new direction is at 60 degrees to the original course, then the traveller would have to push twice as hard ($1/(\cos 60)=2$) to overcome the new reactive force resulting from the direction of the guide wheels. This varying force makes the GuideCane easy to follow, thus helping the traveller to avoid obstacles in their immediate path. Once the GuideCane has made an alteration to the traveller's course, and the object has been passed, the traveller is realigned with their original course by means of an onboard tracking system. In trials it was found that any subject could immediately follow the GuideCane at normal walking speeds around large numbers of obstacles. However, this device is incapable of presenting the user with a representation of their surroundings in any form, except by the number of alterations made to their original direction of travel.

The NavBelt on the other hand, is a portable device equipped with ultrasonic sensors worn on a belt around the waist, and was originally conceived in 1989, and first built and tested in 1994. The belt contains eight sensors covering a combined field of view of 120°. The system has three modes of use, each giving feedback in the form of

binaural auditory cues, based on interaural time difference (phase difference between left and right ears) and amplitude difference to create a *virtual direction*. This is a very useful technique that will be described later in more detail under the heading of ‘Sound outputs’ – section (3.1).

The three modes of use are:

Guidance Mode: In this mode it is required that the system knows the travellers location in respect to their surroundings at all times, as well as their intended destination. In 1998 the prototype system did not contain the required sensors to perform this task. Once operational this mode, using a Global Positioning System (GPS), should allow the traveller to select a destination and the NavBelt would guide them to it. The required travel direction will be indicated to the user by a single stereophonic tone, with the frequency denoting the recommended travel speed.

Image Mode: This mode presents the user with a panoramic acoustic image of their surroundings that is presented via an audio signal that passes from their right ear to the their left. An object can be located by the spatial direction of the signal, with the signal’s pitch and volume corresponding to the object’s distance (higher pitch and volume for shorter distances). It was found that in a cluttered environment this mode presented too much detail for the traveller to comprehend [BorUlr97], so the computer unit suppresses unnecessary scene features [ShoBorKor98]. Only those sections of the environment that are of most importance to the user, such as the closest objects, are signalled in the auditory output.

Directional Guidance Mode: The NavBelt operates in a similar manner to the Guidance Mode explained above, however, in this case the user selects a temporary destination with a joystick (intended to be replaced with speech control at a later date). The destination is located 5m ahead of the traveller in the direction of the joystick. If an obstacle is encountered, then the NavBelt tries to guide the user around it with as little disruption to the intended course as possible.

The NavBelt has been designed so that in an unfamiliar or cluttered environment the auditory display rate increases up to 10 signals per second. In an open environment with few or no obstacles the transmission rate may be as low as one signal every three seconds. This way the user is informed when to be more cautious of their surroundings due to an increase in number of obstacles.

Tests with the prototype NavBelt have revealed promising results. Research presented in a paper by D. D. Clark-Carter, *et al*, in 1986 [ClaHeyHow86] looking into the efficiency and walking speed of the visually impaired using a Sonic PathSounder indicated that a preview range of 3.5m was optimal, and increased walking speeds by 18% over travel with the long cane. [ShoBorKor98]. Similarly, in trials with the NavBelt it was determined that the most favourable results were obtained with a maximum preview range in the region of 3m. Any shorter and the subjects did not have enough time to react to obstacles, any further and the ultrasonic detectors interfered due to crosstalk, which led to unreliable signals.

The average walking speed of sighted people is in the region of 1.3 m/s. Using the NavBelt, after a period of training in the region of 40 hrs, subjects were expected to be suitably proficient to obtain an average walking speed of 0.8 m/s in the Guidance mode, and 0.5 m/s in the Image mode. The reduced walking speed for the Image mode can be explained by the greater complexity of the auditory signal. [ShoBorKor98]

2.3.2. Optophonic systems and mappings

Although the work into the areas of echolocation and sonar-like devices has been shown to demonstrate a reasonable amount of success, they are still greatly limited in terms of their application and use. Furthermore, they all tend to suffer from the same difficulties:

- Multiple sensors (ultrasonic transmitters and detectors) can cause inaccurate readings due to crosstalk [ShoBorKor98]. This is especially noticeable when the maximum range of the sensors are increased beyond 4 or 5 metres, which can also lead to loss of resolution due to ground effects (signal reflection due to the various surface angles found along the ground) [BenBen63].
- Difficulties have often been found to arise through interference from external sources of ultrasonic waves. This tended to be more of a hindrance than a major problem, since this could be overcome by modulating the outgoing signals, thus reducing the problem to one of simply finding the encoded signals in amongst background signals. [Hey85].
- Ultrasonic systems are incapable of detecting visual texture. Due to the nature of echolocation, a page of text or a sign is perceived as a blank surface.
- Similar to the problem above, ultrasonic devices are unable to see through glass, representing a window as a flat surface or wall.
- When approaching a wall at an angle, the wall may not register with the ultrasonic device due to the nature of reflection. A sonar device is dependent on receiving reflected ultrasonic waves as an indication of a nearby object, however, if those waves are being transmitted away from the receiver (i.e. if the incident wave is not normal to the surface of the wall), then the obstacle will not be located. [Mac85]
- Some materials are strongly reflective to certain ultrasonic frequencies, whilst others greatly absorb certain signals, making some objects nearly undetectable with ultrasonic devices.
- Ultrasonics devices also suffer due to the conduction medium – air. Thermal and convection currents can introduce refraction effects that frequently obliterate the signals completely [BenBen63].

In order to solve the above problems, it appears necessary that a blind person have a good, clear representation of their surroundings, as provided by a video camera. The question that arose was whether it would be possible to extend the principle of converting light to sound to that of navigation. A number of people had already thought of this and even progressed so far as patenting devices. These were essentially variations of the original travelling optophone or reading optophone, the main differences being in the scene-to-sound mapping and the output qualities, such as resolution.

The principles behind the video optophone is that a video camera of some form captures images that are generally reduced in resolution and relayed to an electronic device for conversion into sound. This process of conversion from scene-to-sound is called the optophonic mapping. Known attempts at optophonic mappings include:

- **Fish** [Fis76] - 1976

Fish used frequency to map vertical position, and binaural loudness difference for horizontal position, where the sound heard at any instant depended on the brightness gradient of one point in the scene. The scene being represented by a 2-D array of points that was scanned in raster fashion. This form of mapping is classified as a **point mapping**.

- **Dallas** [DalEri80] - 1980

In the patent application, Dallas mapped vertical position in the scene to frequency, horizontal position to time, and brightness to loudness. The mapping segmented a 2-D image into vertical strips that were processed and displayed audibly, scanning from left to right. This form of mapping is an example of the **piano transform**, rediscovered independently by O'Hea [Ohe87] and Meijer [Mei92a].

- **Kurcz** [Kur81] - 1981

Kurcz used another **point mapping**, in the form of a hand-held device called a **heliotrope**, which sensed the light output from only one point in the scene, that point being controlled by the user. The sound output being a function of the light intensity.

- **Deering** [Dee85] **and Tou & Adjouadi** [TouAdj85] - 1985

Deering used a system that incorporated both a tactile display (for directional purposes) and synthesised speech to give an estimated description of any obstacle blocking the user's path. Tou and Adjouadi described a similar system, which informed the user of obstacles in their immediate path, and would try to identify the object from an onboard database. If a match were found the computer would generate a verbal description.

Verbal descriptions of a scene tend to be rather slow and can severely limit the effectiveness of a device. Furthermore, computer programmes are not very effective at visual object recognition, which often requires breaking the scene down to the bare minimum causing a further loss of important features from within the scene.

- **Dewhurst** [Dew99] - 1986 – Present

Dewhurst proposed a patch mapping, giving primarily a sound output, but with the possibility of tactile displays. He considered methods of utilising an area of enhanced resolution equivalent to the fovea centralis in vision.

Areas of particular interest would be presented via a 'language' constructed of consonant-vowel-consonant sounds. This language, used due to the ease of

assimilation and retention by our short term-memory, contains the information corresponding to the structure of the scene.

- **Nielsen, Mahowald and Mead** [NieMahMea87] - 1987

Nielsen, Mahowald and Mead claimed they could enable the perception of motion. They accomplished this by using the time derivative of light log-intensity in a two-dimensional visual field. This was then mapped to an auditory transient (click) filtered so as to appear to come from the same place in a two-dimensional auditory field. In theory, allowing the reconstruction of the third spatial dimension.

As O’Hea [Ohe94] pointed out, this may be so for parallax motion of the camera, however, it is not clear whether the effect is suppressed during panning motion, and if so what is used instead.

Two significant contributions to the study of optophonic mappings in recent years are by Adrian O’Hea [Ohe87, Ohe94] and Peter Meijer [Mei92a, Mei92b]. O’Hea spent 8 years working on his PhD on optophones [Ohea94], titled ‘Optophone Design: Optical-to-Auditory Vision Substitution for the Blind’, at the Open University in the U.K. His thesis is the first major study of optophonics, in which he tried to extract the best methods by comparing other attempts. One main disadvantage at that time was the inability to produce real-time sounds, and so his thesis was largely untested. Unfortunately he was unable to take this work further due to his untimely death shortly after completing his thesis.

The **piano mapping**, which was the same as in the work by Dallas, was considered to be unsatisfactory by O’Hea. For example, conveying a wide light shape on a dark background, the mapping for which is analogous to trying to convey two notes on the piano (the edges of the shape) by playing all the notes in between. A second simulated slot mapping was from edge orientation to musical (circular) pitch, and from position along the slot to interaural intensity difference. He later suggested a

cosine, polar piano and even a free-field patch (the scene is split into segments of interest that are dealt with individually) transform.

He developed a method for evaluating the mappings that he called the Theoretical Performance Test (TPT). This was necessary because it was impractical at that time to carry out trials, as he couldn't generate real-time sounds. The TPT is made up of six stages that make it possible to find flaws in prospective mappings. These stages are:

1. Obtain a digital image.
2. Using the trial-mapping convert the image to sound.
3. Using known inaccuracies of human hearing, calculate an almost perceptibly different sound.
4. Recalculate the scene using an inverse of the mapping.
5. Criticise the recalculated scene visually, by comparison with the original or otherwise.
6. See if any clues emerge to a better mapping.

Although there were a number of limitations with the TPT method, it did enable O'Hea to make a number of extensive evaluations and comparisons between prospective mappings that would otherwise have been impossible without further backing and equipment.

At the same time, Peter Meijer was developing an optophone at the Philips laboratory in Eindhoven. He used the piano transform in a similar way to Dallas, O'Hea and even Fournier d'Albe. Not only has Meijer created a working real-time (software) device, which he has called 'the vOICe', but has also produced an excellent web site [Mei00] that describes his work and provides links to relevant sites. To date this represents the furthest that anyone has gone in optophonics, but he has still to prove its effectiveness as a navigation aid since extensive trials have yet to be carried out.

2.4. The video optophone

2.4.1. Introduction

As previously mentioned the optophone device was modified by a number of parties until its final form, which was one intended for the reading of printed text. It consisted of a row of five selenium cells that would, when passed across a line of printed text, generate different musical notes that corresponded to the shapes of the characters themselves. Using such a device, an extremely adept blind student attained speeds of up to 60 words a minute. However, for most users of the device this speed was far beyond their reach.

Currently, in principle the optophonic mapping still remains the same. Peter Meijer's [Mei92a] modern optophone now uses a video camera to capture a visual representation of the scene, rather than a row of selenium cells, and the sound is computed by small microchips rather than the more bulky circuitry of Fournier d'Albe's day. Furthermore, the optophonic mapping for use in a blind mobility aid now processes a 64x64 pixel display rather than the one pixel of the original exploring optophone.

Peter Meijer used 64x64 pixels for his original video optophone after deriving equation [2.1] that corresponds to the number of bits of information that the auditory system is capable of correctly perceiving due to limitations imposed by crosstalk from the signal. [Mei92a]

$$(M-1).N \leq (B.T)/2 \quad [2.1]$$

Meijer assumed that the communication time, T , should be in the order of one or two seconds, since that corresponds to a period of time that people tend to be most comfortable with in terms of motion and speech. Since the 'useful' bandwidth of the human auditory system, B , is only about 5 or 6kHz (representing the frequencies most readily perceived by people of all ages), and if it is taken that the image dimensions,

M and N, are equal then they are of the order of 50 pixels. But since computers work in magnitudes of two, M and N were set to 64 pixels.

The optophone's video camera captures a digitised graphical representation of the scene, which means that it consists of a series of pixels each having a specific colour or grey level. The image is then broken down into vertical slots, with high frequency (maximum of 5000Hz) musical notes being generated for pixels located at the top of the slot and low frequency (500Hz for the lowest) notes for the bottom. This instantly aids recognition and comprehension of the sounds since our auditory system naturally associates higher frequencies with a higher vertical position in space [Mei92a]. The amplitude of the sounds generated varies proportionally to the intensity of the light captured by the video camera.

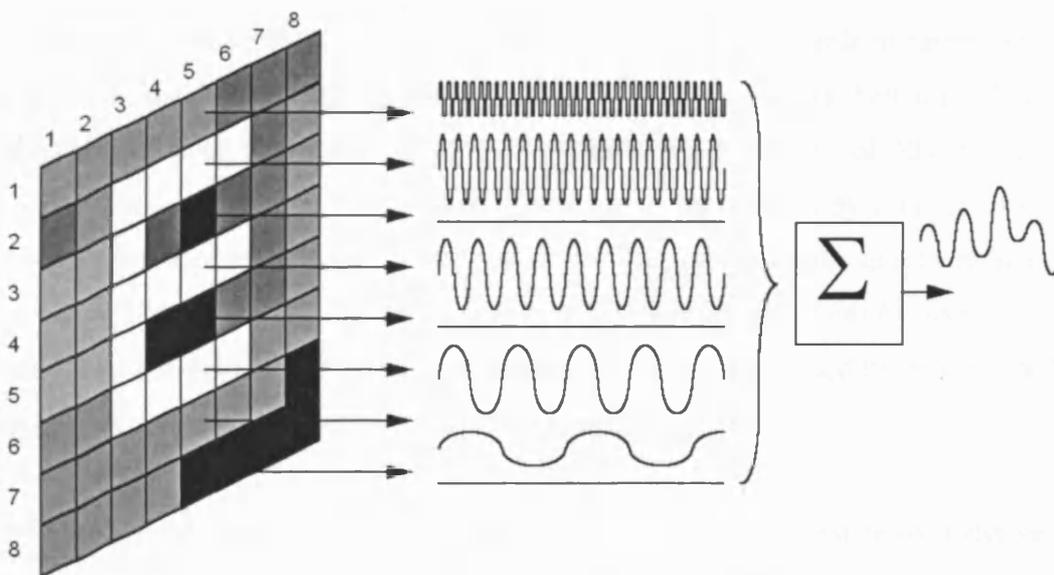


Figure (2.5) – A simplified diagram of Meijer's optophone system

The actual frequency scale used to represent the vertical scan corresponds to an exponential distribution, which better matches the human auditory system and is perceived as being approximately equidistant or linear.

Figure (2.5) demonstrates a simplified version of a transformation from image to sound called a **slot mapping** or the **piano transform**. Aptly named since it can be represented as a vertical piano keyboard scrolling across a 2-D image and generating musical notes as output. The slot is moved across the image from left to right. The sound output generated at any given time is a musical chord solely dependent on the visible portion of the scene (through the slot).

One of the main problems of this method of converting pictures into sounds is that the human ear is nowhere near as good as the human eye for processing data. This means that the information from a picture has to be dramatically reduced. For example, if it is assumed that the video image is a colour VGA image, displayed at 25 frames per second, the number of bits per second is:

$$640 \times 480 \times 24 \times 25 = 184,320,000 \text{ bits per second}$$

The total bandwidth of audible sounds that the human ear is capable of assimilating is about 15 kHz (although this perceived range is often much less, hence the 'useful' bandwidth of 5 or 6kHz mentioned earlier), allowing a bit rate of 30,000 bits per second. This implies that the number of bits has to be reduced by a factor of 6144. One way this can be achieved is to reduce the size of the image and to reduce the number of bits per pixel. For example, a monochrome image of 64x64 pixels is used with 4 bits per pixel and 1 frame per second, providing a reduced bit rate of 16,384 bits per second, which is within the audible range.

Although the optophone may have changed its shape and appearance over the years, there are several factors that have remained the same. The role has remained the same, that of helping the blind in some form. The original optophone was also extremely tiring to use, and required an enormous amount of training and effort from the user to become even close to proficient with the device. This is still the case with the present day optophonic mapping, since the modern versions process at least 4096 (64x64 pixels) times the quantity of pixels for every scan of an image.

2.4.2. Modifications

It was for the reasons described above that the original optophone faded into obscurity, and to a lesser extent its modern day counterpart. Hence, the next aim of the research was to investigate methods for reducing the stress on the user by fading out unimportant regions of the captured images (such that after conversion to sound these regions would have very little emphasis/amplitude), whilst retaining the essential image data required for mobility and avoidance of obstacles.

A possible solution appeared to be available in the form of stereovision, which would provide a method for emphasising important image features (nearby objects), as well as a superior method for indicating the relative distance to objects. In the resulting image, known as a depth map, range would be encoded via the pixel intensity of the image, and thus by the amplitude of the resulting sound. This means that in the optophonic output emphasis would be placed on nearby obstacles by the apparent lack of sound from distant objects. Since avoidance of nearby objects is one of the goals of a navigation aid, this would be appropriate.

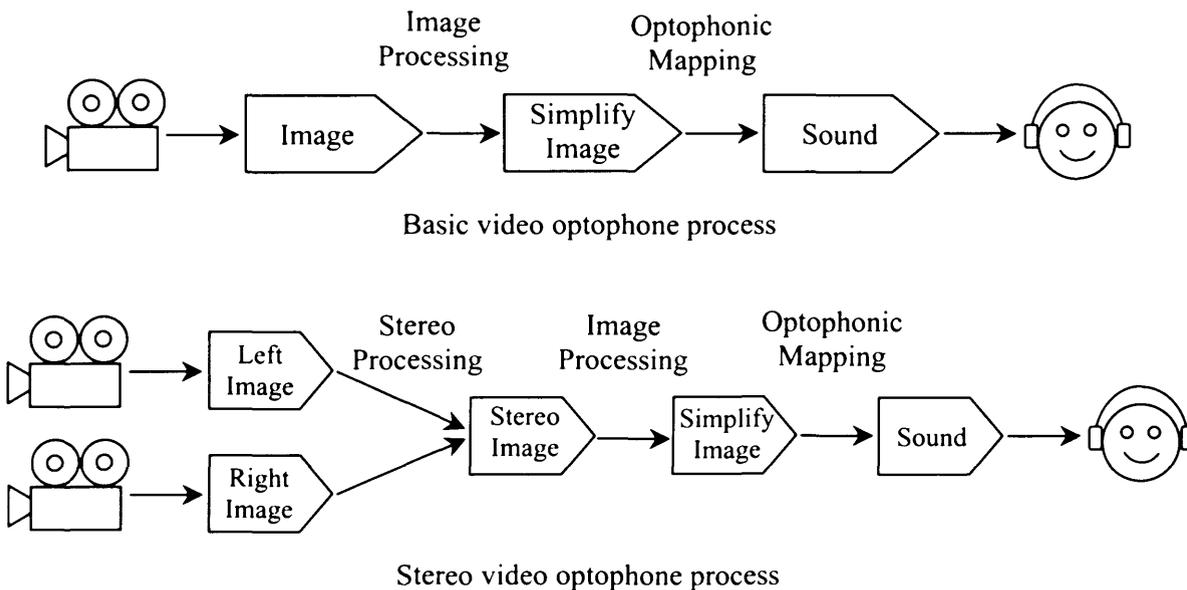


Figure (2.6) – Two diagrams showing the basic processes involved in capturing images and the respective conversions to sound for both the original optophone and the proposed stereo optophone.

Figure (2.6) demonstrates the two described optophonic conversions from scene-to-sound. The first is a more basic system as used in the standard video optophone, which allows for the application of basic image processing techniques as a middle stage in an attempt to reduce the quantity of unimportant features within the captured scene as shown in chapter (3). The second graphical representation incorporates stereovision prior to the more basic image processing stage in an attempt to further improve the emphasis on important image regions, as will be described in greater detail in chapter (4).

Prior to incorporating the proposed systems for stereovision into an optophonic device it was considered necessary to study the more basic forms of scene feature modification, such as edge detection. Firstly, this would help as a basis for the principles of stereo edge processing, which are generally far superior in terms of frame rate than their better looking counterpart, intensity based depth maps. Secondly, if a more simplistic approach proved to be more than adequate for an optophonic device, then there would be less of an urgency to investigate the effects of applying stereo processing.

2.5. Summary

Over the past century there have been numerous attempts at creating the 'ideal' blind aid, both in terms of reading, and later for mobility. The search for the perfect reading aid concluded when technology reached the stage at which OCR and speech synthesis became commercially available. This left only the question of whether a suitable blind mobility aid could be created. Of these devices many used principles of echolocation and ultrasonics, but all tended to suffer from the same complications of interference and general lack of texture. Ultrasonic devices are not the only form of mobility aid that have been built and tested in recent years. There have been a number of optophone variants tried, albeit with only limited success due to the excessive quantities of scene information they portray to the user. In an attempt to improve the link between real world scene and the user's mental image, individuals

have experimented with various image-to-sound mappings. One person in particular, Adrian O’Hea, introduced several very distinct mappings. However, due to the limited computer power at his disposal and restrictions on time, O’Hea was unable to test the usefulness of all of his mappings in a practical sense. Instead O’Hea managed to generate a theoretical method for evaluating the effectiveness of each mapping, which involved encoding an image with the algorithm in question, then decoding it with the inverse of the procedure. By comparing the original and modified images it was possible to make a prediction about the effect it would have on the user, in terms of ease of use. More recently Peter Meijer has even created a software video optophone that is freely available over the Internet.

Although there appears to have been a great deal of progress with video optophone’s in recent years, appearances can often be deceptive. By reviewing the limited results from the various optophone variants one conclusion can immediately be drawn. In each case, the user is unable to make the most of the aid because they cannot fully comprehend the device’s output due to its complexity and high bit rate.

Computers are now sufficiently fast and small enough to be used in a pre-processing role, so that the scene information (i.e., captured images) can be modified prior to presentation to the listener. For basic mobility it can be assumed that a user only needs the essential information, such as the location of obstacles that are in their immediate vicinity, as well as any moving objects that could be on a collision course. If a computer could suppress the surplus features within a scene a user would be provided with a greater opportunity of correctly interpreting the audio output from an optophonic device.

One possible method for suppressing the less important aspects of a scene image would be to generate stereo depth maps. This would effectively reduce the number of objects perceived by the user down to only those within their immediate vicinity. However, before studying stereo techniques it was deemed appropriate to investigate slightly more simplistic forms of image processing.

3. A simplistic approach to information processing & reduction

Various techniques are described for the purpose of reducing the bandwidth of the output from the scene-to-sound mapping used in a software version of a modern day optophone. Also mentioned in some detail are methods for processing the audio output so that as the real world scene is relayed to the user, the location of an object can be determined from the apparent direction of the sound.

Methods of reducing and re-stressing image features that are illustrated in this section include straightforward image reduction and edge detection, as well as a number of procedures for tidying up cluttered images prior to conversion to sound. The techniques demonstrated in this section are all tried and tested methods of information processing, within the realms of sound generation and image processing. These methods were used to guide the research into the area of stereo processing and a further reduction in emphasis of relatively unimportant features contained within the captured images.

As indicated in the previous section some method was required to reduce the complexity of the output relayed to a blind user when using a mobility aid, whilst retaining the essential features to still allow the user to be comfortable in their surroundings. The assumption was made that the tried and tested optophonic mapping was sufficiently suitable and the main problem with optophonic systems was the quantity of information (bit rate) relayed to the user. Hence, there were only two areas that could be modified to any great extent: the sound output, and the graphical representation of the real world scene.

By modifying these two factors to fade out less important features it would be possible to display the output from the optophone in a more pleasing manner, as well as reducing the effects of information overload that the otherwise complex output causes. At the same time retaining sufficient scene structure for the user to

comfortably and easily (ideally) navigate around their local environment. (Not forgetting that the human brain is far superior in filtering image/sound data than any present day computer – so while it may be considered desirable to lessen the sound output, care must be taken not to remove any pertinent detail)

3.1. Sound outputs

An initial improvement was made to the sound output by modifying the blind aid so that it directly assisted the user in locating an object. For instance, it would be much easier to find a real world object if its corresponding sound appeared to come from the same vicinity as the object itself.

By using stereo sound effects it was possible to reintroduce the appearance of direction into the output. Techniques applied include changing the amplitude between the left and right ears depending on the position of the scan. As well as using dichotic presentation, which involves incorporating a shift in the stereo signals of no more than a few milliseconds, effectively emulating the different distance sound has to travel to reach the ears.

The original basic display method used by Peter Meijer's optophone was a monaural sound output, with the horizontal displacement (across the scene) being represented as a function of time and a click indicating the start of each frame. More recently Meijer has improved 'the vOICE' (the software version of the optophone) to incorporate stereo sound, so that the left side of the scene generates sound to be heard (mainly) in the left ear, and the right side of the scene generates sound (mainly) in the right ear. The vertical dimension of the scene being represented by the frequency of the sine waves, since the human auditory system naturally represents (to a certain degree) height with higher frequencies. For example, the human ear recognises (roughly) a logarithmic increase in frequency, as a linear increase in height of sound source. [Fis76, Mei92a]

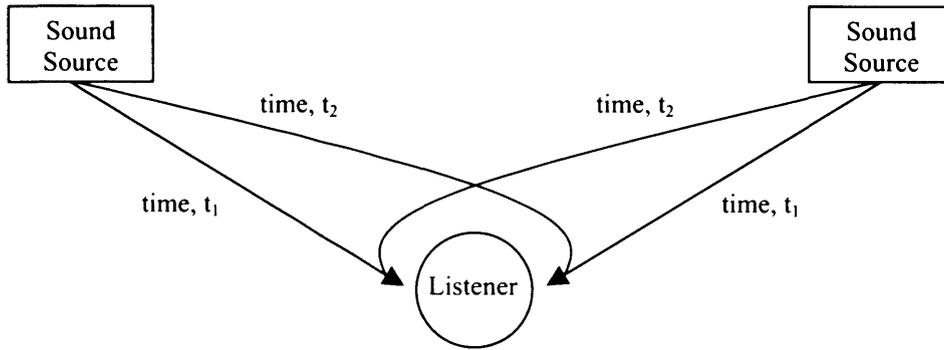


Figure (3.1) – Representation of dichotic presentation, where $t_2 > t_1$ by an order of a few milliseconds.

Past research by E. D. Schubert [Sch56] has demonstrated encouraging results with the use of dichotic presentation (or interaural phase difference) for stereo effects (figure (3.1)). The human brain and auditory system is most adept at ‘tuning in’ on a particular sound source (i.e., a person), even when surrounded by a room full of talking people. One process used to (aid) achieve this is through dichotic presentation. For example, if the sound source is directly in front, then the brain expects the sound to reach both ears simultaneously. However, if the sound source is to the left, then the sound will be heard in the left ear a few milliseconds before it reaches the right ear [Bri98, Hol94, Sch56, ShoBorKor98]. The two sounds are then fused together and used to reinforce each other.

This phase shift can easily be calculated in terms of time difference between the left and right ears [ShoBorKor98], as shown:

$$\Delta t = K_1 \sin \theta \quad [3.1]$$

Where $K_1 = 0.000666$ s is the time phase constant (the distance between the listeners ears, divided by the speed of sound) and θ is the angular position of the virtual source from the plane following the listeners line of sight (figure (3.2)).

The angular shift of a sound source due to interaural amplitude difference is given by equation [3.2]:

$$\theta = K \log \left[\frac{A_R}{A_L} \right] + \theta_0 \quad [3.2]$$

Where A_R and A_L are the amplitudes to the right and left ears, respectively, K is the sensitivity factor, and θ_0 is a constant offset.

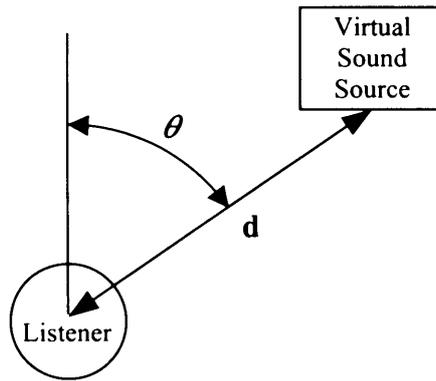


Figure (3.2) – Angular displacement of a virtual sound source.

Shoval, Borenstein & Koren [ShoBorKor98] cite a PhD thesis by D. Rowel [Row70] in which the

sensitivity factor is found to be equal to two for most audible frequencies. Hence, it can be assumed that $K = 2$ and $\theta_0 = 0$. As can be seen from figure (3.1), signals from the virtual sound source will reach one ear before the other. In this case, it is assumed that the amplitude of the sound reaching the listeners first ear is dependent on the distance from the sound source. Consequently, equation [3.2] can be rearranged to provide the amplitude of the sound heard at the second ear.

In this way it is possible to convey an object's location to a listener simply by presenting an auditory signal through a pair of stereo headphones, whilst varying its properties in accordance with the above equations.

Research into signal compression by M. P. Beddoes [Bed68] also demonstrated that modifying an optophonic output to produce physically shorter sounds enabled subjects to learn test signals in a much shorter span of time and to identify them more accurately than they could with the original unmodified set. These 'physically shorter sounds' were achieved by using a signal/time compressor on the output of an optophone like reading machine.

The time compressor takes the original signal, and splits it up into T_s (the sampling interval) and T_d (the discard interval). The final signal is then the smoothed combination of the sampling intervals, taking $T_s/(T_s + T_d)$ of the original time.

For the purposes at hand, the method of signal compression used by Beddoes would be unsuitable. Beddoes' research used an optophone-like device to generate sounds that corresponded to letters of the alphabet. Compressing the signal by simply discarding sections could work since the number of different tones for each letter of the alphabet is rather small. However, for the video optophone under consideration there are 64x64 different regions, some of which would undoubtedly be lost if signal compression were employed to remove sections of the signal. An alternative solution could be to compress the sine wave signal by finding zero-crossings, then applying these to a second smaller arbitrary signal.

For the work described in this project, only the stereo sound techniques were tried, with dichotic presentation being left for future work.

3.2. Image processing

(Note – Many of the techniques discussed in the following sections on image processing can generally be found in the three references [BalBro82, JaiKasSch95, & Low91], unless otherwise specified).

Adrian O'Hea firmly believed that the optophone was the way forward. He argued that an optical input supplies the device with the information available to a sighted person. Echolocation with ultrasonic waves cannot see anything far, through glass, or on paper. Also hearing seems to be able to convey a great variety of information and headphones seem the least invasive form of output device. He also noted that a device with an optical input should not be considered merely as a mobility aid. If a device can report shapes, it can be used for reading, for example, signs and notices in the street. The present research has tried to remain true to O'Hea's beliefs.

It was also believed that a blind user should be provided with the same type of information as a sighted person. An important process that is taken for granted in the everyday life of a sighted person is depth perception. This system is a property of human vision that is provided through the use of two eyes, which supply two images

of any scene, allowing accurate depth calculations to be made. A second process through which the brain perceives depth is via parallax motion that can successfully be achieved with one eye alone through monocular stereovision. For example, by closing one eye and then moving the head in some form of translational motion it is still possible to determine the relative distances to objects. There are other processes involved in the perception of depth, such as the size of an object providing clues about its distance.

The point is that depth perception plays an important role in the everyday life of a sighted person. For this reason the research was directed towards the inclusion of stereovision techniques prior to the scene-to-sound mapping in an optophonic device. Unlike Meijer's original optophone, stereo image processing was included as part of an optophonic system, which will be described later. Firstly, a refresher course was required in the more basic aspects of image processing, also with the intension of studying other possible methods for reducing the complexity of an optophonic sound output.

For the optophone to work as a possible vision replacement system it seems natural that the conversion from scene image to sound be considered in full. In this case the mapping not only needs to reflect the necessary bandwidth reduction when converting between these differing media, but should also be representative of the image processing that our brain performs. Current technology remains inferior to the human brain in terms of the ability to process and filter both visual and audio signals, and thus, it is more reasonable to let the human brain do most of the work, and for the optophone to simply convert the *important* parts of the captured images into another medium – sound. Now, *all that remains to be accomplished* is to decide what features in an image can be classified as important, or what features in an everyday scene does the human brain most readily employ.

In the following sections image-processing techniques are described that can and have been adopted to try to compensate for a possible lack of sight in a user of an optophonic device.

3.2.1. Colour reduction

There are a number of methods that could be used to reduce the workload of the user when attempting to understand an optophone's output. Firstly, a simple reduction in the number of colours/amplitudes that are displayed/heard. For example, a monochrome 256-grey scale image can always be reduced to a 16 level grey scale, effectively reducing the required bandwidth by a factor of 2 from 8-bits to 4-bits. The result is an image with only the greatest contrast changes remaining (figure (3.3b)), which can be considered similar to the human vision system in some respects, since it is adept at detecting edges.

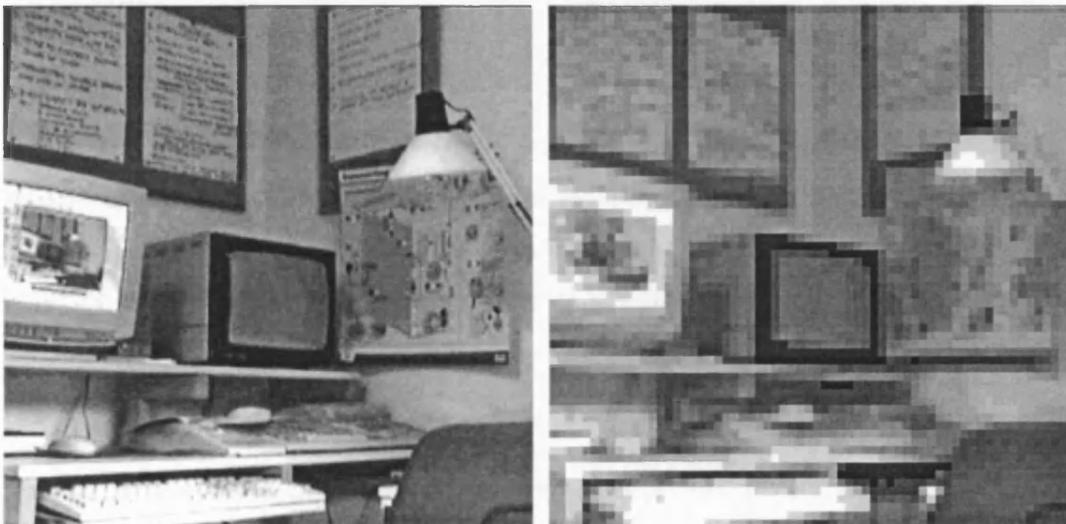


Figure (3.3a) – 256x256-pixel image with 256 grey scales. Figure (3.3b) – Example of the general input to the modern optophone – 64x64 pixels and 16 grey levels.

In a similar manner to that of Dewhurst [Dew99], the workload of the listener may be lessened further by reducing the number of sound levels that can be heard at any one time, by limiting locally the number of different grey levels. Consider an image that contains up to 16 different grey levels, then for each vertical column in the image (since the sound is generated column by column) the four greys of 'best fit' are selected. This can be achieved by finding the positions of the first-, second-, and third-quartiles from the cumulative frequency of the 16 grey levels. Splitting the

colour range into four. The colours of 'best fit' are then taken as the mean (grey level) of each of the four ranges. See figures (3.3a) & (3.3b) for original images, and figures (3.4a) & (3.4b) for the modified images.



Figures (3.4a) & (3.4b) have been generated from figure (3.3b). Figure (3.4a) – The 4 colours of best fit for the whole image. Figure (3.4b) – Still contains up to 16 grey levels; however, each vertical column is composed of only 4 grey levels.

Figure (3.4b), generated by finding the four grey levels of best fit for each vertical column, demonstrates a greater level of visual detail than figure (3.4a), at the cost of small errors that can be seen in the way of stripes down the image.

As seen from figures (3.4a) & (3.4b), applying the colours of best fit does reduce the number of intensities or amplitudes contained within the image and thus, the required bandwidth, whilst retaining enough detail to visually interpret the scene. Whether these images can still be interpreted once they have been converted into sound remains to be ascertained. Although the generated sounds should be less complex than those for figure (3.3b), since for every vertical strip in the image the generated sound would contain only 4 different amplitudes rather than 16.

3.2.2. Noise reduction

Later sections reference various methods of image smoothing commonly used to reduce noise. The reasons for their use will become apparent later in the relevant sections, however, before venturing off into the realms of edge detection and stereo depth maps an endeavour shall be made to briefly discuss some of the more basic methods of noise reduction.

3.2.2.1. Mean filter

One of the simplest low pass filters is called the *mean filter*, equation [3.3]. This works by averaging the pixel intensities, given by $f[k, l]$, from a small square window and replacing the central pixel with the result.

$$g[i, j] = \frac{1}{M} \sum_{(k,l) \in N} f[k, l] \quad [3.3]$$

Where M is the number of pixels in the region N , and $g[i, j]$ is the average for the whole window.

For example, equation [3.4] corresponds to a 3x3 mean filter.

$$g[i, j] = \frac{1}{9} \sum_{(k=i-1) \dots (i+1)} \sum_{(l=j-1) \dots (j+1)} f[k, l] \quad [3.4]$$

Although this type of filter is effective at removing high-frequency components, and hence, removing noise, it results in the loss of much fine detail (figures (3.5c) & (3.5d), page 40).

3.2.2.2. Gaussian filter

The symmetrical 2-D *Gaussian Filter* is very effective at removing noise, and uses a lobe or peak to give weight to the centre of the mask. This central lobe is an improvement over the basic mean filter since it lends more weight to the areas of fine detail, producing a cleaner image (figure (3.5e), page 41).

Equation [3.5] is the two-dimensional Gaussian function. However, there are a number of different approximations to the Gaussian filter, such as the coefficients of the Binomial expansion.

$$G[i, j] = e^{-\frac{(i^2 + j^2)}{2\sigma^2}} \quad [3.5]$$

Where σ determines the spread, or width, of the Gaussian filter. For example, a 5x5 Gaussian filter with $\sigma^2 = 2$ is given in table (3.1), and shown graphically in figure (3.6).

[i, j]	-2	-1	0	1	2
-2	1	2	3	2	1
-1	2	4	5	4	2
0	3	5	7	5	3
1	2	4	5	4	2
2	1	2	3	2	1

Table (3.1) – The above table is the result of applying the Gaussian function, from equation [3.5], with i & j in the range [-2,2] and $\sigma^2 = 2$, followed by multiplying out and rounding to the nearest integer. See figure (3.6) for a graphical representation.

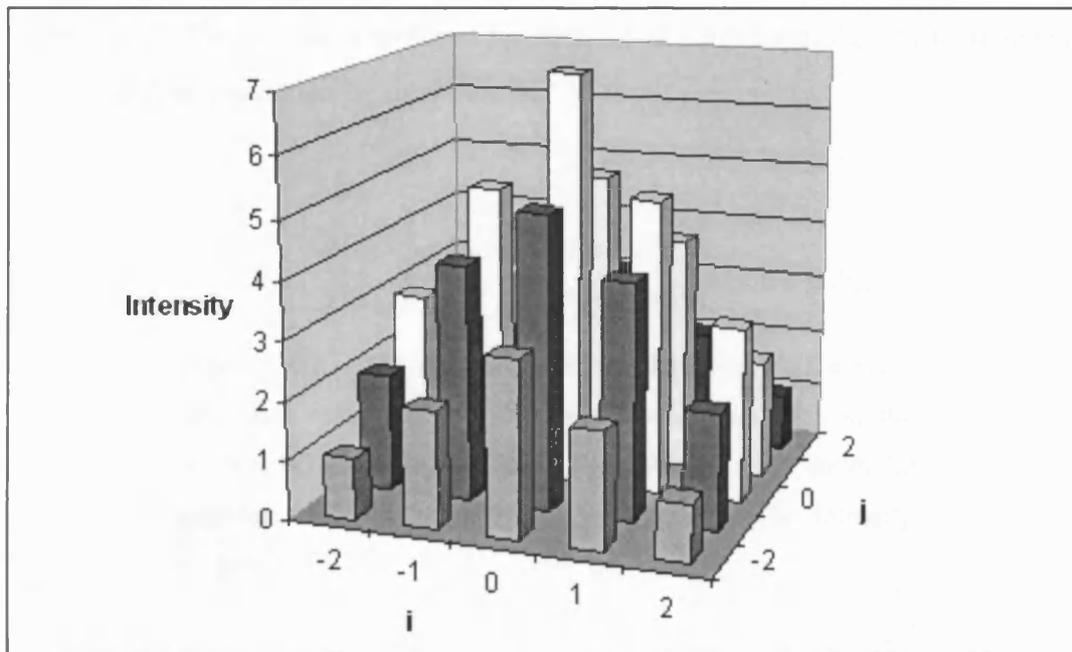


Figure (3.6) – A graphical representation of the 5x5 Gaussian filter with $\sigma^2 = 2$, as defined in table (3.1).

Once the convolution mask has been calculated, as in table (3.1), the image is filtered using equation [3.6].

$$g[i, j] = \frac{1}{M} \sum_{(k,l) \in N} f[k, l].G[k, l] \quad [3.6]$$

Where $g[i, j]$ is the smoothed pixel value, $f[k, l]$ is the image window and $G[k, l]$ the Gaussian convolution mask, both of size N pixels, and M is the sum of all integer weights in the Gaussian mask. For the mask defined in table (3.1), M is 75.

3.2.2.3. Median filter

The median filter is one of the most effective filters for removing random or impulse noise. A square window of pixels is taken from the original image about the point to

be filtered. These are then sorted into order in terms of their grey level intensities, and the middle value chosen (hence the name ‘median’ filter) for substitution into the output image. The process is repeated for each set of pixels until the whole image has been scanned, as illustrated by figure (3.7).

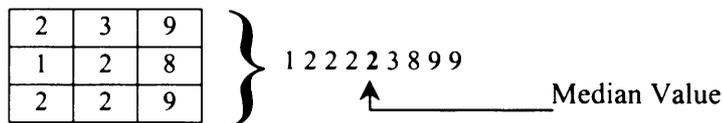


Figure (3.7) – 3x3 neighbourhood median filter. If a mean filter had been used, the outcome would have been 4, the result being a blurring of the edge (the vertical 9 8 9 down the right-hand-side of the neighbourhood, assuming an intensity range of 0-9).

Although the median filter may be very good at removing noise (figures (3.5g) & (3.5h)), it does come at a price. To find the median value, at most $(n/2)+1$ values from the neighbourhood must be sorted, where n is the number of pixels in the neighbourhood. This needs to be completed for every pixel in the image, which, for large images, can be very time consuming, even on a powerful computer.

It can be seen that the central peak in the Gaussian filter (figure (3.5e)), when compared to the mean filter (figure (3.5c)), leaves more of the strong edges (detail) apparently unaffected, while still removing some unwanted fine detail.

Figure (3.5g) & (3.5h) demonstrates just how effective the median filter is at removing fine detail from an image. In some respects the 5x5 median filter may be too effective, and a 3x3 neighbourhood filter would be more suitable. Figure (3.5h) reveals how successful this type of filter is at removing random noise, which is where the median filter excels.



Figure (3.5a) – Office scene. 320x240 pixels and 255 grey levels.



Figure (3.5b) – Figure (3.5a) with 10% added random (pseudo) noise.



Figure (3.5c) – The result of applying a 5x5 mean filter to figure (3.5a).

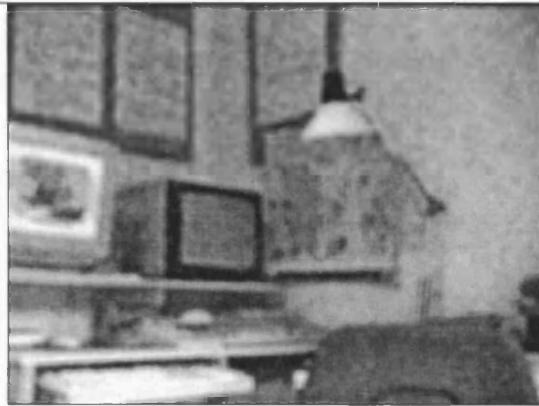


Figure (3.5d) – The result of applying a 5x5 mean filter to figure (3.5b).



Figure (3.5e) – 5x5 Gaussian filter, as defined in table (3.1), applied to figure (3.5a).



Figure (3.5f) – The same Gaussian filter as used for figure (3.5e), applied to figure (3.5b).



Figure (3.5g) – 5x5 median filter applied to figure (3.5a).



Figure (3.5h) – Median filter applied to a noisy image (figure (3.5b)).

There are many other forms of low pass filter, but those described are some of the most common techniques frequently employed. For the following sections any reference to a low pass filter corresponds to a Gaussian filter due to its speed and capability, although any equivalent low pass filter would generally suffice.

3.2.3. Edge & line detection

A more extreme method for reducing the emphasis on unimportant areas of the auditory display would be to incorporate an edge detection routine. The “image” would then consist of cartoon-like images showing the outline of objects. Again this would seem to correspond to the human vision system in its ability to pick out regions of high contrast, or edges.

There are many methods for locating edges in an image, however they can generally be placed into the following categories:

- **Edge detectors**
- **Line detectors**
- **Contour tracers**

Most edge detectors locate pixels that lie on an edge by looking for gradient changes in pixel intensity over a localised region of the image. The gradient, first derivative, is then used as a representation of the edge strength.

Line detectors on the other hand, often incorporate an edge detector, but utilise a method of storing edge information to represent the image in terms of straight lines. The Hough technique [DudHar72, Pic84, Wal85] for instance uses edge angles to transform the image into ‘Hough Space’, whereby all pixels that lie on the same line in (x, y) space are represented by lines that pass through a single point in (m, c) space. Where the equation of a line is given by: $y = mx + c$.

Most contour tracers attempt to locate an edge point then proceed to follow it until either it returns to the starting position, or a predetermined stage has been reached (such as a time limit). The system may proceed with several iterations until the majority of edge points have been traversed. This particular topic shall not be discussed in any depth since these techniques often require a great deal of computation time and the resulting images generally consist of edges of a single pixel

in width, which are unsuitable for the required application (which involves reducing the image dimensions down to 64x64 pixels).

3.2.3.1. Edge detectors

3.2.3.1.1. Gradient operator

Edge detection is the task of finding significant local changes in an image, or rather, finding peak values in the first derivative. If the image is considered as a two-dimensional array of intensity values with function $f(x, y)$, then the gradient is defined as the vector (equation [3.7]):

$$\mathbf{G}[f(x, y)] = \begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{bmatrix} \quad [3.7]$$

The gradient is therefore given, and approximated, by equation [3.8].

$$G[f(x, y)] = \sqrt{G_x^2 + G_y^2} \approx |G_x| + |G_y| \approx \max(|G_x|, |G_y|) \quad [3.8]$$

Therefore, for images, a simple gradient approximation is given by equation [3.9].

$$\begin{aligned} G_x &\cong f[i, j+1] - f[i, j] \\ G_y &\cong f[i, j] - f[i+1, j] \end{aligned} \quad [3.9]$$

Where j corresponds to the x-direction and i corresponds to the negative y-direction.

Note – Using the approximations in equation [3.9], G_x is actually the approximation to the gradient at the interpolated point $[i, j+1/2]$ and G_y at $[i+1/2, j]$. Therefore 2x2 differences are usually used, with the calculated point lying in the centre of the four pixels, as shown in equation [3.10].

$$G_x = \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \quad G_y = \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix} \quad [3.10]$$

Allowing the computation of the gradient of the array $f[i, j]$ in terms of two arrays $P[i, j]$ and $Q[i, j]$ (equation [3.11]) for the x and y partial derivatives (G_x & G_y).

$$\begin{aligned} P[i, j] &\approx (f[i, j+1] - f[i, j] + f[i+1, j+1] - f[i+1, j])/2 \\ Q[i, j] &\approx (f[i, j] - f[i+1, j] + f[i, j+1] - f[i+1, j+1])/2 \end{aligned} \quad [3.11]$$

The magnitude ($G[f(i, j)]$) and orientation of the gradient is then defined as:

$$\begin{aligned} M[i, j] &= \sqrt{P[i, j]^2 + Q[i, j]^2} \\ \theta[i, j] &= \arctan(Q[i, j], P[i, j]) \end{aligned} \quad [3.12]$$

This technique is a simple edge gradient detector, with the gradient magnitude being used as an indication of the edge strength. A drawback being that the resulting edge information does not correspond to any pixel in the original image. Instead, the edge magnitude relates to an imaginary point lying between the four pixels of the filters in equation [3.10]. To avoid this problem the Sobel Operator is often used.

3.2.3.1.2. Sobel & Prewitt operators

The Sobel operator uses a 3x3 neighbourhood to calculate the gradient with subsequent edge information corresponding to the centre pixel and not an interpolated point.

The Sobel and Prewitt operators are given by the magnitude of the gradient computed by equation [3.13].

$$M = \sqrt{s_x^2 + s_y^2} \quad [3.13]$$

Consider the arrangement of pixels about the pixel $[i, j]$, shown in figure (3.8).

a_0	A_1	a_2
a_7	$[i, j]$	A_3
a_6	A_5	A_4

Figure (3.8) – The arrangement of pixels under consideration when using the 3x3 Sobel operator.

Then the partial derivatives can be defined as:

$$\begin{aligned} s_x &= (a_2 + ca_3 + a_4) - (a_0 + ca_7 + a_6) \\ s_y &= (a_0 + ca_1 + a_2) - (a_6 + ca_5 + a_4) \end{aligned} \quad [3.14]$$

With the constant $c = 2$ for the Sobel operator, or 1 for the Prewitt. Convolution masks for the Sobel gradient operators are shown in figure (3.9).

-1	0	1
-2	0	2
-1	0	1

s_x

1	2	1
0	0	0
-1	-2	-1

s_y

Figure (3.9) – Convolution masks for the Sobel edge operator.

The above edge operators all compute the first derivative, and if above a certain threshold, the corresponding pixel is assumed to be part of an edge (figure (3.10b)). Hence, they usually generate wide edges, which can be a failing when pixel accuracy is required.

3.2.3.1.3. Laplacian operator

To rectify this problem, the second derivative can be taken, and an edge point set when a zero crossing is encountered. Since the zero crossing does not rely on any threshold values for edge classification, it can produce very thin and accurate lines (figures (3.10c) & (3.11)).

The Laplacian is the 2-D equivalent of the second derivative given by equation [3.15].

$$\nabla^2 f(x, y) = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \quad [3.15]$$

From equation [3.9] the following difference equations can be obtained, which are used to approximate the second derivatives along the x and y directions:

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= \frac{\partial G_x}{\partial x} = \frac{\partial(f[i, j+1] - f[i, j])}{\partial x} = f[i, j+2] - 2f[i, j+1] + f[i, j] \\ \frac{\partial^2 f}{\partial y^2} &= \frac{\partial G_y}{\partial y} = \frac{\partial(f[i, j] - f[i+1, j])}{\partial y} = f[i+2, j] - 2f[i+1, j] + f[i, j] \end{aligned} \quad [3.16]$$

These approximations are centred about $[i, j+1]$ and $[i+1, j]$, respectively. Therefore, they can be re-centred giving:

$$\begin{aligned} \frac{\partial^2 f}{\partial x^2} &= f[i, j+1] - 2f[i, j] + f[i, j-1] \\ \frac{\partial^2 f}{\partial y^2} &= f[i+1, j] - 2f[i, j] + f[i-1, j] \end{aligned} \quad [3.17]$$

Combining these two equations gives the operator shown in equation [3.18]:

$$\nabla^2 \approx \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad [3.18]$$

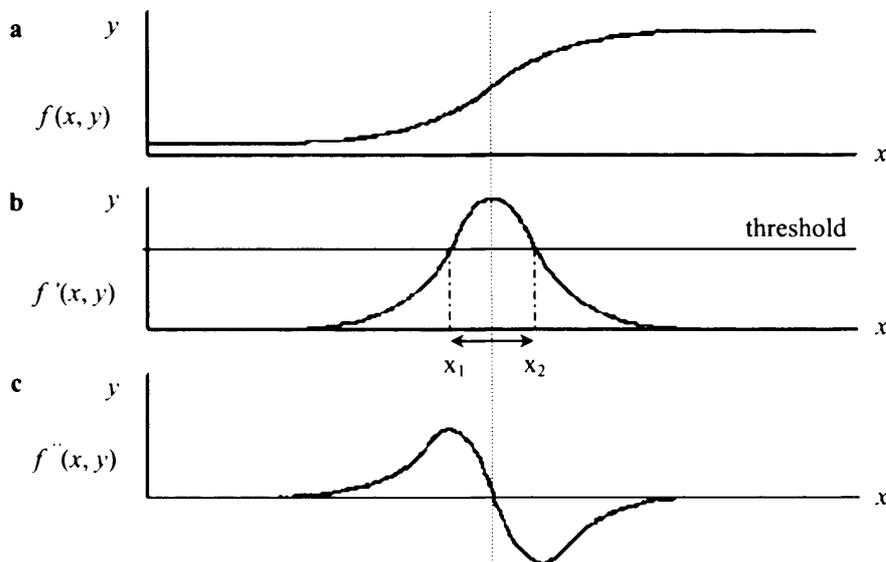


Figure (3.10) – (a) Edge in the image function $f(x, y)$. (b) First derivative of the image function. Wide edges are formed in the region above a pre-set threshold between x_1 and x_2 . (c) The zero crossing of the second derivative provides distinct edge maxima.

As demonstrated in figures (3.10c) & (3.11), the Laplacian operator identifies an edge when the output makes a transition through zero.

(The output is given by: $F = \sum_{i=1}^3 \sum_{j=1}^3 P_{ij} C_{ij}$, where F is the filtered value of the target pixel, P is a pixel in the 2-D grid, and C is a coefficient in the 2-D Laplacian matrix)

target pixel, P is a pixel in the 2-D grid, and C is a coefficient in the 2-D Laplacian matrix)

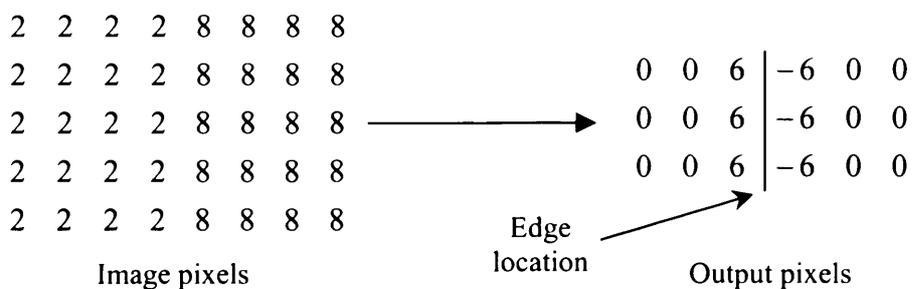


Figure (3.11) – Application of the Laplacian operator to an image edge. Once a result has been obtained, interpolation is used to find the location of the true zero crossing.

However, the second derivative compounds a number of problems over the first derivative. For instance, the first derivative is inclined to introduce a certain amount of noise, so undoubtedly, the second derivative only adds to the problem. In fact, since it is difficult to incorporate a threshold into the Laplacian operator, even the smallest peak in the first derivative will cause a zero crossing in the second. Another drawback of the operator comes from the lack of directional edge information, the usefulness of which will be made apparent later in the section on stereo edge detection (section 4.2.2).

The most significant problem was realised when considering the application for which the Laplacian operator was intended. The edge operator would be applied to a relatively large image, and then reduced in size to 64x64 pixels. Any thin edges would tend to vanish in the image reduction stage, indicating the need for an edge operator that generated thick edges, rather than one that created thin (one pixel wide) edges, albeit highly accurate.

It is possible to limit the effect of some of these problems. For example, a simple method of reducing noise in an edge image is to apply a low pass filter prior to edge detection. The *Laplacian of Gaussian* (LoG) does just that [Hil82, MarHil80]. It incorporates a Gaussian low pass filter in the process of edge detection, indicating the presence of an edge when a large peak is found in the first derivative, and a zero crossing in the second. Similarly, directional information can be obtained by using the *second directional derivative*, which is the second derivative computed in the direction of the gradient. However, the problem of significant loss of detail during image reduction still remains.

It is possible to solve the problem of narrow edges prior to image reduction by using an edge thickening routine. However, it seemed pointless to waste more computation time on that form of procedure when edge operators already exist, such as the Sobel operator, that readily provide thicker edges.

As can be seen from figures (3.12c) & (3.12e), the 3x3 Sobel operator is less susceptible to noise than the 2x2 gradient operator. In general, the larger the filter

used, the less susceptible it is to noise, but the greater the quantity of fine detail that is lost and the higher the price in terms of computation time. Conversely, the smaller the filter used, the greater the noise, but the better the detail and speed of operation.

An alternative approach is through the application of a low pass filter, such as the 5x5 Gaussian filter used for figure (3.12b), which has much the same effect as increasing the size of the edge operator. A low pass filter effectively smoothes out the rapid intensity changes that often correspond to noise, whilst leaving the stronger edges. However, some fine detail is often lost, as is the case with the background text in figures (3.12b), (3.12d) & (3.12f), and especially figure (3.12h).

The Laplacian filter produces very fine and accurate edges, however without any pre-filtering the quantity of unwanted detail is too great (figure (3.12g)), and with filtering too much of the required detail is removed (figure (3.12h)).

Thus, the Sobel operator appears to be a suitable algorithm for the generation of edges for the application at hand. It can not only provide thick edges ready for the image reduction stage of the optophone (prior to conversion to sound), but it also provides more accurate edge information [Pic84, Pic89] than many detectors, such as the edge orientation, which is useful for generating stereo edge depth maps – section (4.2.2).



Figure (3.12a) – Office scene. Note - In the background fine detail exists in the form of text.

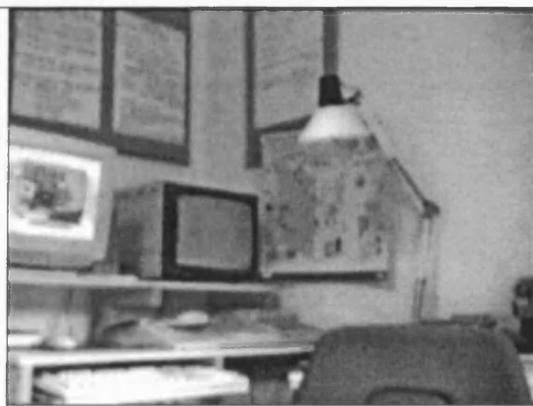


Figure (3.12b) – 5x5 Gaussian filter. Some fine detail (i.e. background text) and noise has been removed.



Figure (3.12c) – Gradient edge operator (equation [3.11] & [3.12]). For clarity and for the sake of comparison all edges above threshold (of 10) were set to white (255).



Figure (3.12d) – Gradient operator applied to figure (3.12b).



Figure (3.12e) – Sobel edge operator (equation [3.13] & [3.14]) with properties as for figure (3.12c).

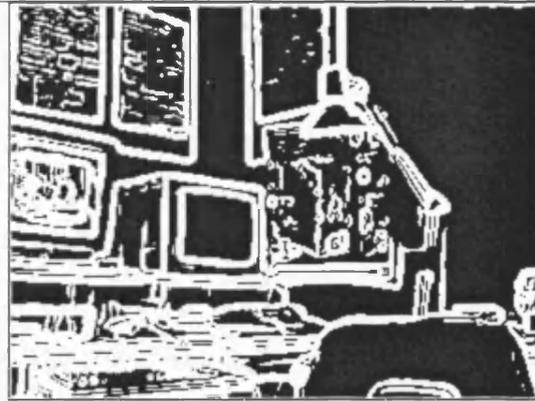


Figure (3.12f) – Sobel operator applied to figure (3.12b).

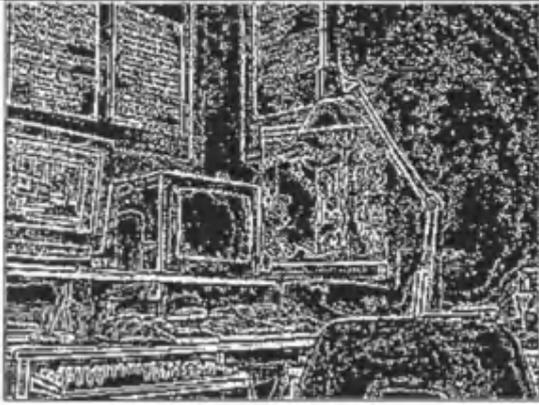


Figure (3.12g) – Laplacian filter, equation [3.18], with edge threshold set to 10.

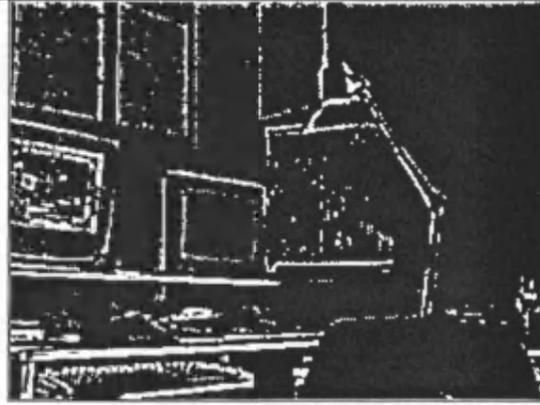


Figure (3.12h) – Laplacian filter applied to figure (3.12b).

3.2.3.2. Line detectors

There are numerous techniques currently used for detecting and storing line data [BurHanRis86, DudHar72, Pic84, Wal85]. One of the most commonly used is the Hough Transform [DudHar72, Pic84, Wal85], which works by transforming edge/line data obtained from the equation of a line, into a two-dimensional space that groups pixels lying on the same line into peaks. Consequently, strong lines from the original image can be located by performing peak detection within the Hough space.

For example, the general equation for a line is, $y = mx + c$, thus it follows that every line passing through a point (x_i, y_i) must obey, $m = (y_i - c)/x_i$. All points in the Hough domain that obey this equation will be incremented, and so any peaks found in the Hough domain correspond to a line in the initial image (figures (3.13a) & (3.13b)).

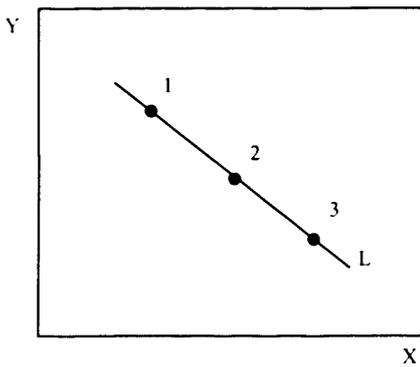


Figure (3.13a)

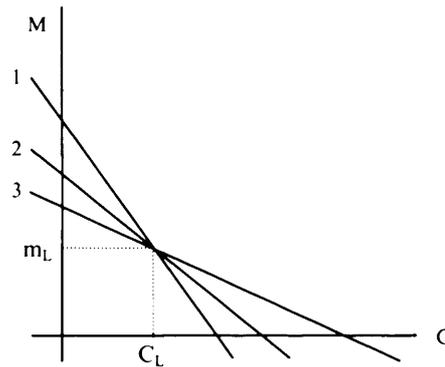


Figure (3.13b)

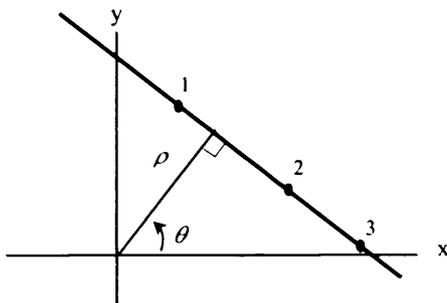


Figure (3.13c)

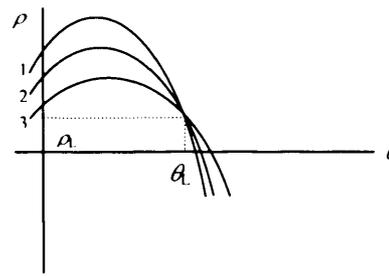


Figure (3.13d)

Figure (3.13a) – Image line passing through three defined image points. Figure (3.13b) – Classical Hough transform into m - c space, representing image line of figure (3.13a) as a convergence of points. Figure (3.13c) – Transformation of edge coordinates into ρ – θ coordinates with the range, $-\text{Image_Size} \leq \rho \leq \text{Image_Size}$ and $0 \leq \theta < \pi$. Figure (3.13d) – Modified Hough space; whereby real world edges are represented by convergence of lines formed by edge points after transformation into ρ – θ coordinates. Labels 1, 2, & 3, define edge points in figures (3.13a) & (3.13c) and correspond to respective lines in figures (3.13b) & (3.13d), after transformation.

However, a problem presents itself when applying this technique to a computer. The values of m and c can exist anywhere within the range of minus to plus infinity. Thus, Duda & Hart [DudHar72, see also Pic84] solved the problem by utilising a different set of parameters that have a restricted range: ρ (the orthogonal distance between the line and the origin) and θ (the angle of the slope minus $\pi/2$). See equation [3.19] and figures (3.13a) & (3.13c).

$$\rho = x_i \cos \theta + y_i \sin \theta \quad [3.19]$$

A common technique for limiting the quantity of detected peaks in the Hough space is through the use of a predefined edge threshold to remove weak edges. Predefined thresholds are often found to be rather inconvenient and restrictive. An alternative is to use the edge strength when incrementing cells in the Hough domain. Rather than incrementing a cell by one each time an edge pixel is located, the cell is incremented by a value that is proportional to the pixel's edge strength. Thus a strong edge or line will generate a very strong peak in the Hough domain, whereas a weak edge will produce almost no peak. [Pic84]

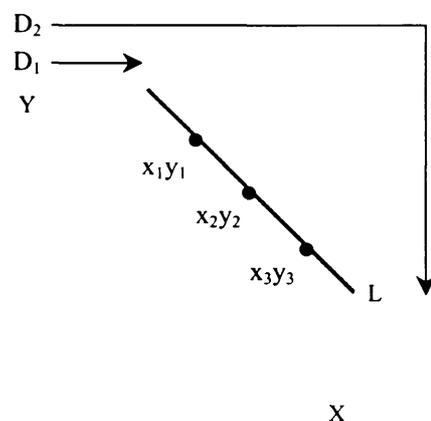


Figure (3.14) – Modified Hough domain, similar to the Muff domain proposed by Wallace.

To speedup the computation time (and to reduce the array sizes to conserve memory space) a possible quantisation that could be imposed with the $\rho-\theta$ transform would be to take lines at every k degrees. However, if a feature were to lie on a slope between the quantised angles, then a cluster would form in the Hough domain, which would be flat-topped rather than peaked. This would only be acceptable for purposes of approximation or if a sophisticated cluster detector were used to find the centre of gravity.

A slightly different parameterisation was proposed by Picton [Pic84], and later a similar technique known as the Muff transform by Wallace [Wal85], both of which defined a line by the points at which it intersected the perimeter of the image. This of

course means that four parameters are needed to characterise a line – the x and y values for the two perimeter points. Although this can be reduced to two if the points are converted to the distance from the origin to the points of intersection, which are represented as D_1 & D_2 in figure (3.14). This technique is more accurate at positioning lines since its coordinates are defined by the perimeter of the image. Whereas the previous techniques are dependent on an angle or gradient, consequently the positioning of coordinates and lines become less accurate further from the origin.

A drawback with the Hough technique for generating edge images comes from its inability to correctly classify two separate edges that lie along the same image line. Unless care is taken, a single peak is formed in the Hough domain resulting in one line rather than two individuals (figures (3.15a) & (3.15b)).

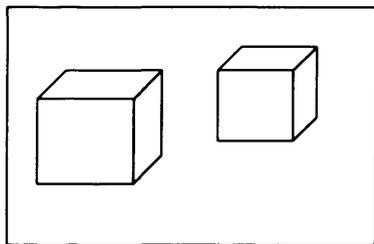


Figure (3.15a)

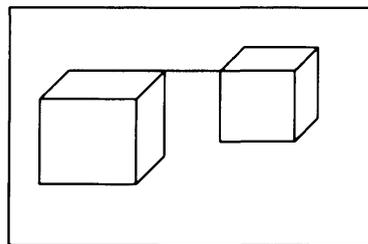


Figure (3.15b)

Figure (3.15a) – Simple scene with two objects. Figure (3.15b) – Expected result of the Hough transform to figure (3.15a). Two of the edges of the cubes lie along the same image scanline; consequently the Hough transform links the edges.

Techniques such as the Hough transform tend to provide very little information about the lines that they attempt to describe. The best one can expect are parameters such as the start and end points of the line, the number of pixels in the line, and occasionally contrast information. Consequently, the Hough transform appears limited in its scope for expansion (for example – its role for stereo matching). Thus, Burns *et al.* developed a technique that attempts to group and build line support regions, providing line information including the average intensity, number of pixels, orientation, average width, contrast, steepness and straightness of the line.

In its simplest form the process of building line support regions contains only a few steps. Firstly, an edge detection routine is used to locate strong edges, and to determine the corresponding edge orientation. Secondly, a partitioning scheme is chosen to quantise the angles into similar regions (i.e., 45-degree segments). Edge pixels are then labelled depending on the partition in which their orientation lies. Finally, a connected-components algorithm is employed to link the support regions that are identically labelled, and to allow for the calculation of a line that best fits the connected region. To further improve the accuracy of building the line-support regions, and to prevent fragmentation of line regions, Burns *et al.*, suggested the use of two overlapping partition schemes, offset by half a region (i.e., 22.5-degrees). The chosen line would consist of the pixels that provided the longest support region obtained from the two partition schemes. This solved the problem of edge pixels that were visually part of a line, but consisted of angles that crossed partition boundaries, and consequently, would not be linked.

Techniques such as those described here for generating representations of a scene in terms of straight lines are generally very effective. However, they often take considerable amounts of processing time making them impractical for a near real-time system, as required for the video optophone. Much important detail can also be lost, such as text in a scene, due to errors or inefficiencies in the edge pixel linking process (figures (3.16b) & (3.16d)). Methods do exist that try to improve accuracy, although these greatly increase computation time. Furthermore, the edges generated by the Hough line detector are too thin for use as input to an optophonic mapping without performing edge thickening, requiring further processing.

One possible use for the Hough line detector that was considered later on in the research was for use in a stereovision algorithm that employed a single camera – monocular stereovision (section 6.4.3).



Figure (3.16a) – Image with dimensions 320x240 pixels, and an intensity level ranging from 0-255.



Figure (3.16b) – Hough lines formed from figure (3.16a). Edge thinning and Hough space peak thinning employed to limit detected lines. A line was declared valid if it had at least three edge pixels supporting it.

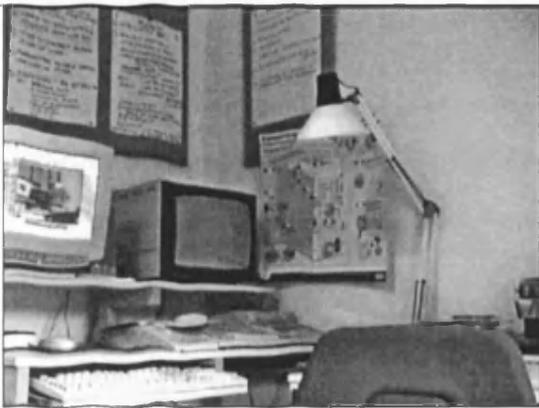


Figure (3.16c) – Office scene.



Figure (3.16d) – Hough image generated from figure (3.16c).

3.2.4. Time varying edge detection

Another important aspect of the human vision system is its ability to perceive moving objects, such as a bird flying past. For this reason it was considered important that the role of motion detection for use in an optophonic blind mobility aid be investigated. Motion detection would, in most cases, greatly reduce the complexity of the

optophone's output since sound would only be relayed to the user when there was something occurring (moving) in the scene.

A basic method for finding moving objects is to simply take the difference in intensity of corresponding pixels between two consecutive frames, or alternatively, by subtracting the images [JaiKasSch95, Low91] (figure (3.17c)). Although, as indicated by this image, it is sometimes difficult to correctly identify a moving object from the resulting motion image due to blending as has occurred with the screen of the laptop. In figure (3.17c) the screen of the laptop appears transparent, revealing the wall behind. To avoid this problem a more common approach locates moving edges, displaying a moving object by its outline alone. This form of detector is known as a 'time varying edge detector', since it locates edges that have changed positions during the time frame that corresponds to the two captured images. [HayJai83, JaiKasSch95, Pic89]

A generally accepted technique uses the product of the Sobel edge operator and the intensity difference between frames [HayJai83, Pic89]. A moving edge being characterised by an edge that is both bold and moving rapidly with respect to time, as given by equation [3.20].

$$\text{Strength of moving edge} = D(\text{frame2, frame1}) \cdot S(\text{frame2}) \quad [3.20]$$

Where S corresponds to the Sobel edge operator, and D represents the absolute intensity difference between corresponding pixels from two consecutive frames. Finally, a threshold is set so that only the boldest edges remain (figure (3.17d)).

There is an obvious drawback with this approach when applied to a blind mobility aid. The user, whilst motionless, will only perceive moving objects in the scene. To detect other objects, the user would have to move the camera, which would result in every edge in the captured images being perceived as moving. The result of which is analogous to that of a standard edge detector, as shown in figure (3.17g).

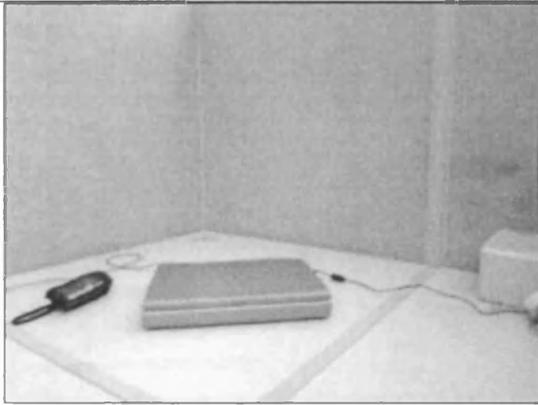


Figure (3.17a) – Frame 1 of a pair showing a laptop computer and a mobile phone.

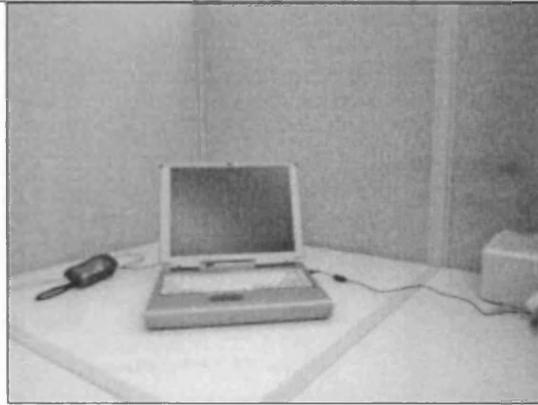


Figure (3.17b) – Frame 2 that shows the same laptop and mobile phone after some obvious movement.

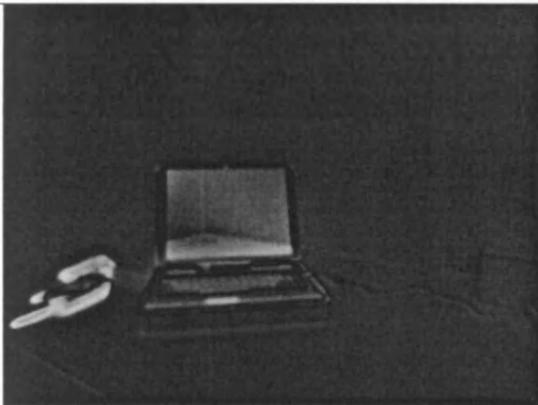


Figure (3.17c) – The result of subtracting the two frames shown above.

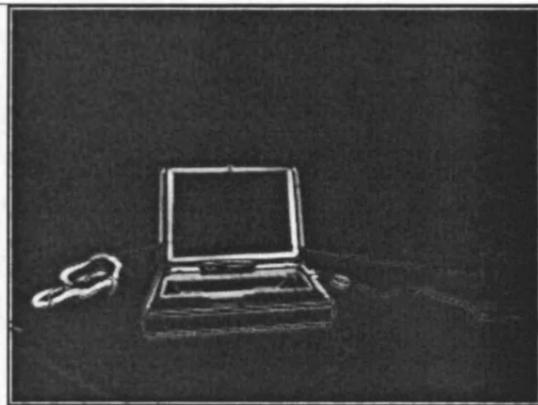


Figure (3.17d) – Moving edges as found by the detector in equation [3.20].



Figure (3.17e) – Image one of a pair with dimensions 320x240 pixels, and an intensity level ranging from 0-255.

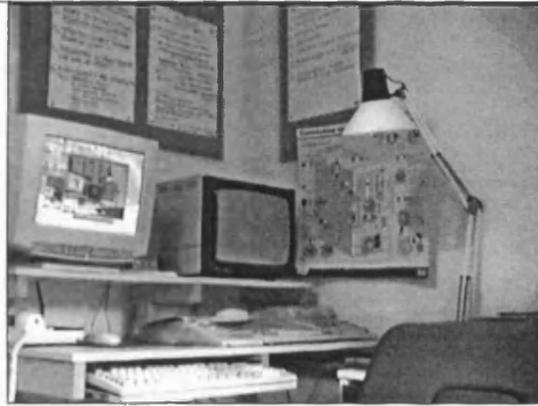


Figure (3.17f) – Image two of a pair with properties as for figure (3.17e).

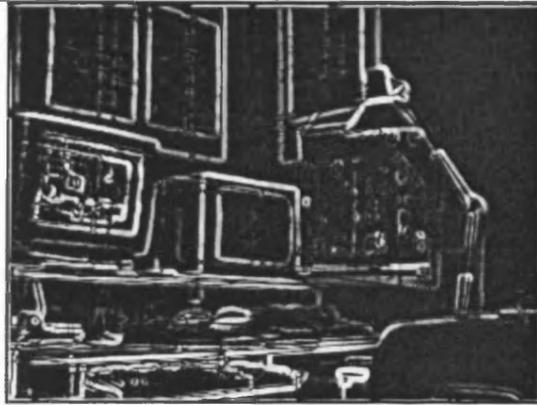


Figure (3.17g) – Time-varying edge detector applied to frames (3.17e) & (3.17f). All edges are visible since a horizontal camera displacement occurred between image frames.

3.3. Summary

As has been demonstrated, many of the techniques described can be employed to reduce or re-emphasise features within an image in an attempt to make the output easier to comprehend. However, none of these procedures preserve depth information that would readily enable a person, blind or otherwise, to gauge the relative distance to any object in the local environment. Furthermore, many of the above techniques actually appear to remove or conceal an excessive amount of important detail that is necessary for basic mobility. For example, many of the methods of calculating edge

boundaries are insufficient for visually determining the true nature of objects in the scenes with which they are attempting to portray. Thus, if it is visually demanding to comprehend the modified images, then it cannot be expected that the task be any less difficult for an optophone user after the alterations caused by the optophonic conversion from scene-to-sound.

Although the individual techniques described in this section may be insufficient for use with an optophonic mapping, when used as a basis for stereo processing they could still provide adequate results.

4. Stereo processing of images

Multiple frames of a scene enable the generation of stereo depth maps that present an image such that important regions are emphasised, whilst retaining depth information. Distance to an object is thus encoded via the brightness of image pixels, and similarly, through the amplitude of the sound output.

Techniques discussed include intensity and edge based depth maps using multiple cameras. Due to limitations encountered during the research into these methods of stereo processing the discussion not only presents the most common techniques, but also continues by demonstrating two original concepts for generating depth maps. These concepts revolve around the principles of generating superior approaches to the presentation of images. The first is a new approach for generating extremely fast edge depth maps, whilst the other demonstrates a new display form for a depth map by incorporating a technique known as cartooning. This method of stereo cartooning, unlike most intensity based depth maps, is fast enough to process frames at real-time speeds whilst retaining areas of shading unlike edge based depth maps.

In the previous section multiple frames were used to demonstrate how to detect time-varying edges. Multiple frames can also be employed in the creation of stereo depth maps, which could be used to provide a blind optophone user with the perception of range information. Also mentioned in the last chapter (section (3.2)) was Adrian O’Hea’s belief that the image-to-sound mapping employed by the optophone should mimic the human vision system as closely as possible in terms of visual perception. Since the optophone attempts to partially replace a person’s lost sense of vision then a suitable way in which to do this is by replicating the visual abilities of the human vision system. An important factor that seems to have been overlooked on many occasions during past research into optophonic devices was the human vision system’s ability to perceive depth. By incorporating depth into the image processing stage of the optophone two important goals were achievable. The emulation of the

human vision system, and a vast reduction in the complexity of the optophone's output.

Before considering the use of a stereo algorithm to generate an image ready for conversion to sound, it was necessary to determine what image or sound property should be sacrificed to make way for the depth information. As with most depth maps it seemed reasonable to relate distance in the real world scene to the relative pixel intensity in the depth map, and to the amplitude of the generated sound, since with the standard video optophone, amplitude corresponds to texture or colour (neither of which are vital to basic mobility). Thus, a distant object beyond the range of a blind user, and of no immediate importance for basic mobility, would be discarded from the output. Alternatively, the distant object would simply produce a much fainter sound than one closer at hand. This substitution of qualities deprived the system (and user) of the ability to detect texture. But this limitation was put aside in favour of the more desirable qualities of the depth map.

It is important to realise that the purpose for investigating various stereo techniques during this research was not to invent new and more advanced stereo algorithms, but rather to study ways in which previous procedures could be moulded to suit the information processing stage prior to the application of the optophonic mapping. A suitable stereo algorithm would not need to generate a perfect depth map, and its resolution could remain quite low considering the final image prior to conversion to sound would only be 64x64 pixels in size. Thus, an ideal stereo algorithm should be capable of generating reasonable depth maps at high frame rates. More importantly, it should be able to represent the scene in such a way so as to strengthen regions of significant interest (objects that are in the user's immediate path), at the same time minimising the distraction to the user from relatively unimportant regions of the scene.

This same principle has also been used when deciding upon the best procedures for use with the optophonic mapping. If the depth maps resulting from a stereo algorithm do not visually portray scene information that can be considered relevant for improved mobility, then that algorithm should not be used. In many cases the simplest and oldest techniques best suit the required criteria.

4.1. Background to stereovision

Stereo depth maps are commonly generated from two or more image frames of a scene, which demonstrate a translational displacement between each frame, as would be provided if a line of cameras were used. This displacement provides a perspective with which to determine object distances within the images. For two cameras with optical axes aligned and a baseline oriented in the horizontal direction with the world origin located midway between the cameras, then the 3-D co-ordinates may be computed directly from the disparity of corresponding image points (equations [4.1]-[4.3]):

$$x = b \frac{(x'_l + x'_r) / 2}{x'_l - x'_r} \quad [4.1]$$

$$y = b \frac{(y'_l + y'_r) / 2}{x'_l - x'_r} \quad [4.2]$$

$$z = b \frac{f}{x'_l - x'_r} \quad [4.3]$$

Where b is the baseline length, f is the focal length, and x'_l and x'_r are the x-coordinates in the left and right images respectively, where the disparity is equal to: $(x'_l - x'_r)$.

From this it is possible to calculate an object's real-world distance with respect to the cameras, as long as corresponding image points can be located. The process of pixel or region matching is achieved by finding a distinct set of features in one image, which can be used to identify likely candidate matches in the second. This process would take a considerable amount of time if the line of displacement (epipolar line) between the cameras were unknown, since the process of finding a match for a feature in the first image would necessitate a search of every feature in the second image. Fortunately, if the epipolar line is known the task becomes much simpler.

If the cameras are correctly aligned (with camera scan-lines lying along identical horizontal rows) then the epipolar lines for both cameras will be along corresponding image rows (figure (4.1)). This is known as the epipolar constraint, which is useful for two reasons. Not only does it indicate what image line to search on in the second image for a match, but also in what direction to proceed, demonstrated in figure (4.1). Consequently, the number of candidate matches is greatly reduced, with the whole task being completed more accurately and efficiently.

Assuming there are n features to be matched with an upper disparity limit of d pixels, then without the epipolar constraint there are $n(2d)^2$ or $4n(d)^2$ comparisons that must be made to find the best match. Using the epipolar constraint as described above, the search region can be reduced to a line containing just nd pixels.

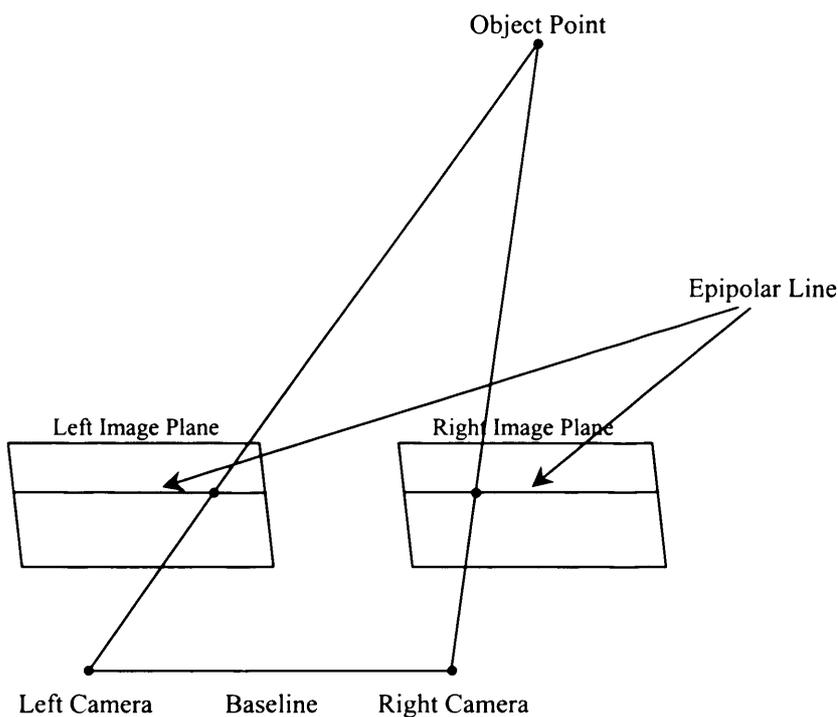


Figure (4.1) – Epipolar constraint for a pair of stereo cameras.

Once a match is found and believed correct, then the disparity is taken and converted into its corresponding pixel intensity using the relation given in equation [4.3]. This range value is then entered into the subsequent location in the final depth map. Equation [4.3] indicates that the disparity is inversely proportional to z , depth in the

real world, and hence, the pixel intensity. However, a linear relation between disparity and pixel intensity appears to provide a more suitable range of depths for use with the optophonic mapping.

4.2. Stereo techniques

Over the past couple of decades many stereo techniques have been derived that can generate a depth map from a sequence of images [Ber97, Fua93, GenMor76, JaiKasSch95, MarDurCha85, MciMut88, Mor96, PolMayFri85]. Amongst these techniques there are generally only two categories of depth maps, intensity based and edge based.

However, a few techniques exist that do not fit into these two categories, such as depth from defocusing or refocusing one or more images of a scene taken with differing lens settings. For example, a frame is captured of a scene with the focus set so that only the nearest objects are clear. The range of real-world objects can then be determined by gauging the amount by which they are out of focus. [GeiKla94, KlaGeiBov95, RajCha97]. For more information see section (6.4.4).

A few methods have also been developed that are aimed at enhancing the output quality of the depth map. These usually incorporate different constraints into the matching process, such as the PMF algorithm [PolMayFri85] that uses the cyclopean disparity gradient (section (4.2.3)).

4.2.1. Intensity depth maps

Intensity depth maps are created using an area-based approach. Correspondences are found by searching the stereo images for regions of identical or similar texture. An ideal intensity depth map would be completely filled with varying pixel shades, with the pixel intensity being proportional to real-world depth. However, perfect depth maps are rarely generated due to occlusions (caused by overlapping objects) or a lack

of texture for matching. Some techniques try to estimate these areas by comparing with adjacent disparities; alternatively uncertain areas may be marked as unknown [Fua93].

A common approach, when generating intensity based depth maps, is to pass a correlation mask or filter over a pair of stereo images looking for the best match [Fua93, GenMor76, JaiKasSch95, Mor96]. This can be achieved by first selecting a point in one image, then for all candidate matches in the second image, the regions, known as windows, surrounding the two chosen points are compared. For instance, this may consist of finding the sum of the squared differences of the intensity values between corresponding points in the two windows, as in equation [4.4].

$$\sum_{[i,j] \in R} (f - g)^2 \quad [4.4]$$

Where R is the region of the window, and $f[i, j]$ and $g[i, j]$ are the stereo images. This can be expanded to give equation [4.5].

$$\sum_{[i,j] \in R} (f - g)^2 = \sum_{[i,j] \in R} f^2 + \sum_{[i,j] \in R} g^2 - 2 \sum_{[i,j] \in R} fg \quad [4.5]$$

If the light intensity of both images is alike, then it may be assumed that both the sum of f^2 and g^2 are nearly constant. In this case the sum of fg (used in generating figures (4.2c) & (4.2d), page 72) will provide a suitable value for the mismatch between candidate points, with a maximum indicating a likely match. However, there are a number of drawbacks with this technique, including the inability to locate matches if any rotation occurred between image frames, and inaccuracies due to the two stereo images having differing average light intensities. Hence, Gennery & Moravec [GenMor76] suggested two different approaches – *normalised correlation* and *pseudo-normalised correlation*.

- **Normalised correlation**

The RMS average displacement of co-ordinate pairs $(A[i, j], B[i, j])$ from a least squares regression line, relating image A to image B . Let $a[i, j] = A[i, j] - \bar{A}$ and $b[i, j] = B[i, j] - \bar{B}$, where \bar{A} and \bar{B} are the means over the entire window, which has dimensions $i=0$ to n , and $j=0$ to n . The normalised correlation coefficient can then be defined as shown in equation [4.6] [Fua93].

$$\sigma = \frac{\sum_{i=0}^n \sum_{j=0}^n a[i, j] b[i, j]}{\sqrt{\left(\sum_{i=0}^n \sum_{j=0}^n (a[i, j])^2 \sum_{i=0}^n \sum_{j=0}^n (b[i, j])^2 \right)}} \quad [4.6]$$

With $\sigma = 1$ for a perfect match between the two windows. Unlike the sum of squared errors, normalised correlation is unaffected by linear variations in illumination between the two windows, although special care must be taken with uniformly shaded regions of an image, since σ will tend to infinity as $a[i, j]$ or $b[i, j]$ tends to zero.

- **Pseudo-normalised correlation**

Gennery & Moravec [GenMor76] proposed a second, improved correlation method (equation [4.7]).

$$\sigma = \frac{2 \sum_{i=0}^n \sum_{j=0}^n a[i, j] b[i, j]}{\sum_{i=0}^n \sum_{j=0}^n (a[i, j])^2 \sum_{i=0}^n \sum_{j=0}^n (b[i, j])^2} \quad [4.7]$$

Although the pseudo-normalised correlation method is more efficient in terms of computation time since it does away with calculating the square root, there tends to be little difference in terms of correct correspondences between the techniques of

equations [4.6] & [4.7]. Figures (4.2e) & (4.2f), page 73, demonstrate the use of the pseudo-normalised correlation method for two different window sizes.

Once a suitable stereo technique has been decided upon, methods can be employed in an attempt to help reduce the number of mismatches encountered. One such procedure utilises a two-way correlation search [Fua93, MciMut88]. This process can greatly reduce the number of incorrect matches since the same false match is rarely made in both directions when performing a two-way search between stereo images. The process finds all possible correspondences between a pair of stereo images, A and B, firstly from image A to image B, then from B to A. The depth map is generated based on the criteria that a true match exists if and only if a correspondence appears in both lists. By ignoring uncertain matches, the depth map itself becomes an indication of the reliability of the matches. If all isolated matches are also rejected then the likelihood is that all remaining matches will be correct. Consequently, a dense depth map suggests that the disparities are reliable, whereas a sparse depth map implies the opposite. However, the improvement this approach affords comes with an overall reduction in speed.

The final depth map may also be improved by increasing the number of input stereo images – whenever possible three or more images should be used. In this case, the first image of a sequence, the reference image, is processed with all other images in turn. When finished, the resulting disparity maps are combined so that matches are reinforced, and uncertain matches are ignored. The combined depth map that this generates is often purer than any individual stereo image. Alternatively, the stereo algorithm can be employed on two images, and then when any candidate match is encountered it can be verified in a third (n^{th}) image.

Some researchers (e.g., Pascal Fua [Fua93]) have taken the approach of combining the results from several different sized correlators, which are individually passed over pre-Gaussian smoothed stereo images. The separate depth maps are then combined, in the same manner as above, providing an improvement in quality. Likewise, Marr & Hildreth [MarHil80] used a similar approach to accurately determine the location of image edges. The process they employed used an edge operator with a range of varying filter sizes to generate a superior edge image. Edges would only be accepted

as valid if at least two filters agreed on their location. However, even though the improvement can be quite noticeable, the processes involved are rather computationally expensive. For this reason, a pyramidal approach [Fua93, JaiKasSch95] can be employed to produce similar results in a shorter computation time. Instead of increasing the filter size at each stage, the image dimensions are reduced, producing the desired outcome and reducing the computation time. Nevertheless, if a correlation filter is being repeatedly passed over the stereo images, the whole process cannot be expected to run in real-time on a standard computer (unless the system is being applied to *very low-resolution* images).

Similarly, it is possible to visually enhance a depth map by combining the outputs generated from slightly misaligned stereo images [Fua93]. Due to the finite resolution of captured images, real-world features are often positioned at slightly differing coordinates than those expected. For instance, between a pair of horizontally aligned cameras, a feature's vertical placement may differ by one or two pixels between the captured images. An improved depth map can therefore be obtained by combining the results given by scans that correspond to the correct image alignment, and those given by a vertical misalignment of one or two pixels. This method also helps correct errors caused by camera misalignment or poorly calibrated stereo images [Fua93].

4.2.2. Edge depth maps

Using any of the stereo procedures previously described it is possible with only two adjacent image frames of an area and a known epipolar line, to obtain a good graphical representation of the scene, whereby pixel intensity is proportional to object distance.

The limitation of this form of depth map comes as a result of its quality. An intensity based depth map can provide a reasonable estimate on the distance of an object from the observing cameras as long as sufficient textural information is available. In most cases this means that the whole image will be processed in the search for similar regions of texture (candidate matches). Indicating the relationship between depth map quality and computation time. The higher the quality, the greater the computation

time, often in excess of several minutes on a standard computer for large high resolution depth maps. Which, considering the need for rapid processing from image-to-sound, is inadequate.

At the loss of some image detail a considerable speed increase can be achieved by altering the search parameters from every pixel in the image to only those of particular interest, such as those on an edge boundary or corner [GenMor76, HarSte88, JaiKasSch95, KorZim86, Mor96]. For example, if edge detection is incorporated prior to the correlation process, then only edge points need be compared, so considerably reducing the number of comparisons [Ber97, JaiKasSch95, MciMut88]. For this reason, the investigation was directed towards the possible use of some form of stereo edge depth map. Although these can often be visually inferior due to their restriction to regions of rapidly changing contrast, or edge boundaries, they have the advantage that they can be processed at fairly high frame rates (multiple frames per second).

Typically, correspondences for stereo edge detection are made by the comparison of edge pixel properties, such as edge magnitude, orientation and changes in contrast (positive or negative). The Sobel edge detector is often used for these purposes due to its accuracy in determining edge orientation.

As an example, edge features can be determined by applying the Canny edge detector as shown in the following algorithm:

1. *Smooth the image with a Gaussian filter.* (To reduce the number of weak edges before edge detection)
2. *Compute the gradient magnitude and orientation using finite-difference approximations for the partial derivatives.*
3. *Apply non-maxima suppression to the gradient magnitude.*
4. *Use the double thresholding algorithm to detect and link edges.*

Both steps 1 & 2 were described earlier (section 3.2.2.2 and 3.2.3.1.1 respectively). The simple gradient edge operator used in step 2 can quite easily be replaced with the Sobel edge operator for greater accuracy of edge features. Step four is of course unnecessary, since it relates to locating and expanding edges to form edge contours. Similarly, step three is only necessary if edge thinning is required prior to edge matching. The process of edge thinning, or non-maxima suppression, is a simple case of locating the true edges in the magnitude array. This is accomplished by thinning the broad edges in the magnitude array so that only the magnitudes at the points of greatest local change remain.

The depth map resulting from the use of step 3 would consist of lines mainly of one pixel in width, which for the purposes of image reduction prior to sound generation would be unsuitable. However, the process of applying non-maxima suppression is quite a simple one and does have a couple of advantages. Firstly, since the correlation stage of stereo processing is normally very computationally expensive, having fewer edges for matching lowers the burden of calculating correspondences. Secondly, if there are fewer edge pixels, then there are generally fewer incorrect matches made during the matching process.

A second form of stereo edge depth map, or rather line depth map, that can be generated uses a line detector, such as the Hough technique, to locate and then match lines. However, there are a number of problems with this method of stereo matching. These include the long computation time that line detectors generally have, as well as the difficulties that can occur when many lines in the stereo images have similar orientations, since they all become likely candidates for each other, and so mistakes are often made.

For more information on line matching see section (6.4.3).

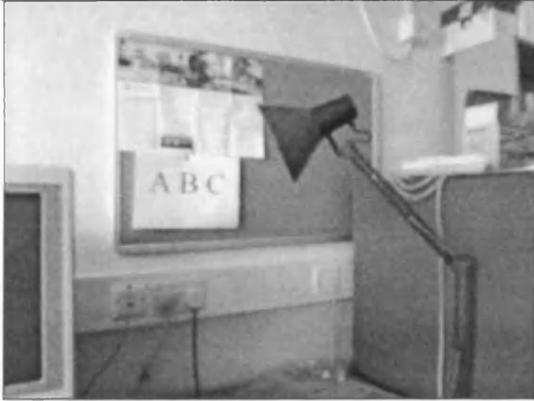


Figure (4.2a) – Image one of stereo pair. Dimensions 320x240 pixels, and a pixel intensity of 0-255.

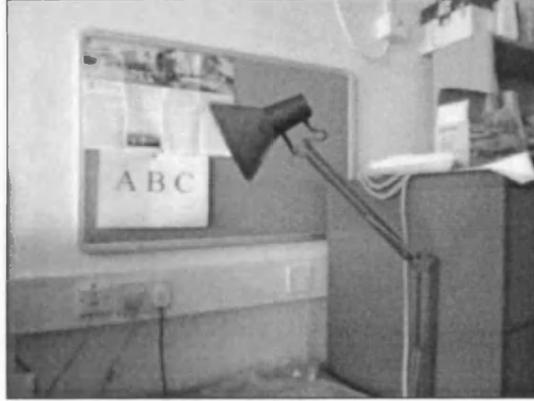


Figure (4.2b) – Image two of stereo pair.



Figure (4.2c) – Intensity based depth map generated using sum of squared differences (equation [4.4]) with a 5x5-pixel search window and a disparity range of 14-30 pixels.



Figure (4.2d) – Sum of squared differences with a 9x9-pixel correlation window.



Figure (4.2e) – Pseudo-normalised correlation given by equation [4.7], using a 5x5-pixel correlation window.



Figure (4.2f) – Pseudo-normalised correlation using a 9x9-pixel correlation window.

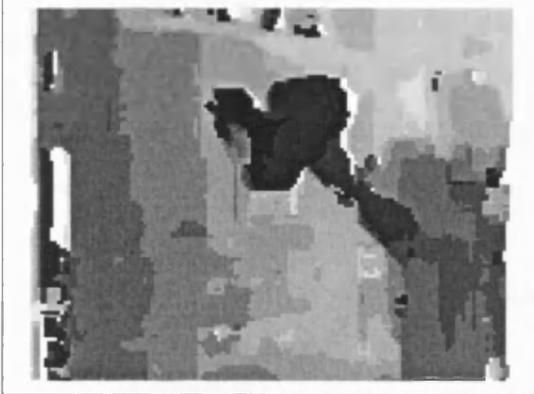


Figure (4.2g) – A 9x9 pseudo-normalised correlation performed upon 160x120 pixel input images.

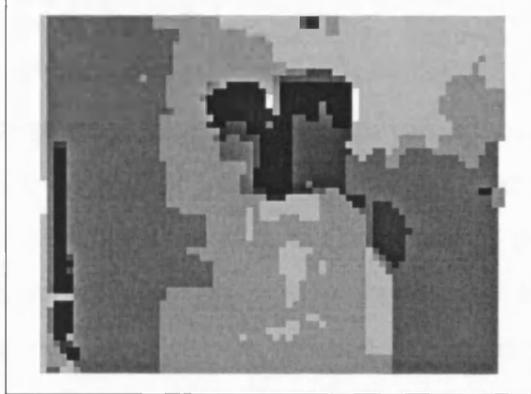


Figure (4.2h) – 9x9 pseudo-normalised correlation performed upon 80x60 pixel input images.

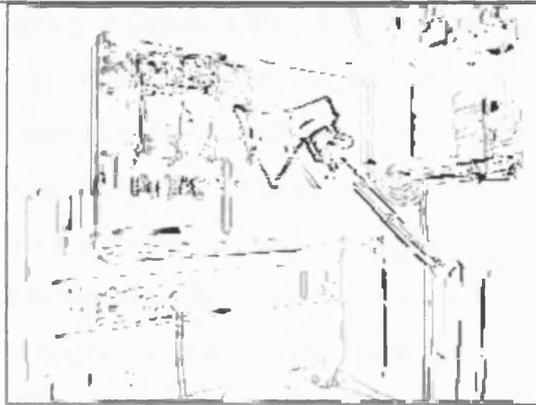


Figure (4.2i) – Edge depth map derived from Sobel edge images of figures (4.2a) & (4.2b). A pixel was declared an edge if its edge magnitude was ten or greater.

The sum of squared differences (equation [4.4]), used when generating figures (4.2c) & (4.2d), suffers greatly when searching for correspondences between areas of little texture. In these instances the resulting pixel intensities, which represent the depth, are often inaccurate. For example, part of the filing cabinet at the bottom right of figures (4.2c) & (4.2d) is displayed in black, implying a nearness to the camera that is not the case. It should also be noted that the increase in correlation window size does little to remedy this problem. This is not the case for the pseudo-normalised correlation (equation [4.7]) demonstrated in figures (4.2e) & (4.2f). Although the problem caused by lack of texture is still apparent as speckled areas that lack any uniform intensity, increasing the filter size greatly reduces the problem (figure (4.2f)).

Figures (4.2g) & (4.2h) demonstrate the effect that using smaller input images has on the resulting depth images. Figure (4.2h) is approximately the same size as that used for input to the video optophone. It can be seen that although much of the fine detail has been lost, it is still possible to perceive objects within the scene, including their approximate depth. Due to the reduction in image size the overall computation time is greatly reduced, unfortunately it is still insufficient (with current computers/PC's) to process more than one frame every couple of seconds when applying the optophonic mapping.

The edge depth map shown in figure (4.2i) maybe far from perfect, but the technique is capable of processing numerous frames per second on a standard PC, unlike the intensity based depth maps previously described. Using this form of edge depth map it is also possible to identify object boundaries, and to determine some aspect of their distance from the viewing cameras. Unfortunately there are obvious errors, such as the black vertical lines around the filing cabinet. This is due to the stereo correlation routine matching image edges incorrectly. The black lines in question result from the correlator incorrectly selecting the first match it finds as correct, where the true match lies further apart.

4.2.3. Matching constraints

A common method for improving the depth maps obtained from a particular stereo technique is to incorporate additional matching constraints. In the stereo edge algorithm described in section (4.2.2) the only constraints used were the edge magnitude, orientation and contrast. If the edge pixels being compared differed greatly in terms of those three criteria, then the pixels would be deemed an unsuitable match. Although these three simple conditions often work, they are still very basic and fail in some circumstances.

A well-known technique used to enhance the quality of a depth map is the PMF algorithm [PolMayFri85]. This uses a value defined as the disparity gradient, which for the correct fusion of two stereo scenes should not exceed a certain limit for all matching points. The PMF algorithm makes the assumption that the disparity gradients between correct matches will be small almost everywhere. Pollard, Mayhew and Frisby [PolMayFri85] found that with a camera geometry approximating the arrangement of the human eyes, a disparity gradient limit of 1 would almost always be satisfied between correct matches arising from a large class of the surfaces forming the visual world. This is not true of incorrect matches.

Consider figure (4.3), a simple stereogram made up of two dots (**A** and **B**) in each field (left and right). When A_L is matched to A_R , and similarly for **B**, the disparity gradient between them is the difference in their disparities divided by their cyclopean

separations. The cyclopean separation, as defined in equation [4.8], is given by the distance between the midpoints of the two pairs of dots (located at A_c and B_c respectively).

$$S = \left\{ \left[\frac{1}{2}(x + x') \right]^2 + y^2 \right\}^{\frac{1}{2}} \quad [4.8]$$

As the change in disparity between the two matches is $x' - x$, then the cyclopean disparity gradient between A_c and B_c is given by equation [4.9].

$$\Gamma_D = |x' - x| \left[\frac{1}{4}(x + x')^2 + y^2 \right]^{-\frac{1}{2}} \quad [4.9]$$

For both dots to be binocularly fused simultaneously by the human visual system, or similar camera arrangement, the ratio of the disparity differences between the dots to their cyclopean separation must not exceed a limit of 1.

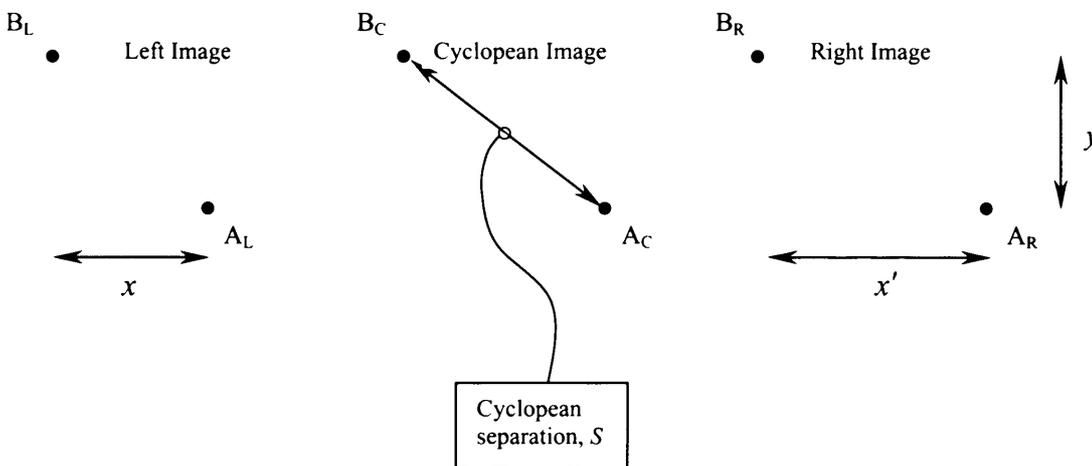


Figure (4.3) – Determining the cyclopean separation for two matching points within a stereogram.

One apparent drawback of using this technique is that it can fail to find matches when a rapid change in depth occurs across the scene. In this situation the cyclopean disparity gradient may exceed the chosen limit and no correct matches found. Fortunately, incorrect matches are rarely made since these also exceed the limit.

4.3. Development of stereo techniques

As can be seen from figures (4.2a)-(4.2i), by using various methods of stereo processing it is possible to reduce the apparent complexity of the optophonic output by fading out distant objects within a captured scene, whilst emphasising those points that lie within the observer's immediate vicinity, and hence those that can be considered of greater significance.

The following section details the equipment, systems and algorithms used for generating real-time stereo images, along with the developments and modifications implemented to the techniques previously described.

4.3.1. Triclops camera system

The camera system used for testing the stereo algorithms was the Triclops system from Point Grey Research in Canada [PGR00], which consists of three cameras in an L-shape configuration. The captured images are then fed directly into a Matrox Meteor Frame Grabber for processing. Each of the three cameras returns image data via a separate colour channel (Red, Green, and Blue). In this way it is possible to capture images at very high frame rates. By combining the high frame rate from the Triclops cameras with efficient pre-processing techniques and stereo algorithms it is possible to generate intensity based depth maps and edge based depth maps at speeds of up to 15 frames per second on a 450MHz PC. (See appendix 3 for information on the hardware and software used, as well as problems encountered during the research and various methods of software optimisation).

The Triclops system incorporates its own software library, which provides suitable functions for the processing of intensity based depth maps. However, modifications have been made during the application of new stereo techniques [CapPic00b], whilst attempting to find the most suitable image representations for the required feature reduction prior to the scene-to-sound mapping.

4.3.2. Stereo depth maps

The intensity based depth maps that are an innate part of the Triclops software library employ the Sum of Absolute Differences, similar to equation [4.4] described earlier, to locate correspondences. For example:

$$\min_{d=d_{\min}}^{d_{\max}} \sum_{i=-\frac{m}{2}}^{\frac{m}{2}} \sum_{j=-\frac{m}{2}}^{\frac{m}{2}} |I_{\text{right}}[x+i][y+i] - I_{\text{left}}[x+i+d][y+i]| \quad [4.10]$$

Where d_{\min} and d_{\max} are the minimum and maximum disparities, m is the correlation mask size, and I_{right} and I_{left} are the right and left images.

The third camera is applied to verify the existence of a possible match, greatly reducing the quantity of incorrect correspondences with very little extra overhead in terms of computation time.

The Triclops software library is split into two processing blocks that consist of pre-processing and stereo processing. The pre-processing allows for low-pass filtering to help prevent aliasing whilst applying rectification to the raw camera images. The rectification compensates for any camera lens distortions and misalignments that might otherwise cause inaccuracies. A further option is the use of edge detection, which allows matching on the changes in brightness rather than the absolute values of image pixels. This option tends to be useful only when the auto gains of the individual cameras do not change simultaneously by an identical amount, which can cause some mismatches due to variations in the brightness of like regions between the stereo images.

The stereo processing stage also accommodates for texture and uniqueness validation whilst searching for correspondences. Texture validation simply determines whether the correlation window under consideration contains enough texture to provide a reliable feature match. Similarly, uniqueness validation can be applied to ascertain

whether the best match for a particular pixel is significantly better than the alternatives. If not, it is likely that the true match is not visible due to an occlusion. If either of these validation methods fail, then the pixel is declared invalid.



Figure (4.4a) – Top Triclops image of a set of three. Figures (4.2a) & (4.2b) are the left and right images, respectively, which makeup the remaining two images. Image dimensions - 320x240.



Figure (4.4b) – Intensity based depth map generated from Triclops images (figures (4.2a), (4.2b), & (4.4a)). Triclops correlation – sum of absolute differences (equation [4.10]) – with 5x5-pixel search window and disparity range of 14-30 pixels.



Figure (4.4c) – Intensity based depth map with a 9x9-pixel correlation window.

As can be observed from figures (4.4b) & (4.4c) the Triclops system produces adequate results, and at a sufficiently high frame rate. It is also clear that if the disparity range was chosen correctly it would be possible to reduce the complexity of the optophonic output by fading out distant objects within the captured images, which after conversion to sound would be portrayed with a less distracting lower amplitude. In most cases this would be advantageous since distant objects generally have less influence on the blind traveller. However, it is not currently possible to create a fast system that would automatically calibrate the disparity range to suit all environments, without additional assistance from the user. This means that the depth map may not, for many real-world scenes, sufficiently reduce the complexity of the final output. For this reason, research continued by studying ways in which the various types of depth map, such as edge based depth maps, could be used to further emphasise important features within a real world scene whilst fading out unwanted detail. A secondary target, and a further reason for investigating edge depth maps, was to include some textural and similarly textual information in the optophonic system. For example, providing the blind user with the ability to perceive and hopefully recognise large text (as in the form of signposts or warning signs – ‘EXIT’ signs), which is not possible with standard intensity depth maps.

4.3.3. Stereo edge depth maps

The stereo edge algorithm described earlier utilises the pre-processing stage of the Triclops software library, with the addition of a feature detection routine that provides edge pixel information such as orientation and magnitude. This is accomplished with the Canny or Sobel edge operator. [JaiKasSch95]

The second stage, stereo processing, uses a separate set of functions that analyses the edge data from two cameras, searching for possible matches. This is accomplished by searching for candidates between the image pair that have near identical edge properties within a given disparity range. Another constraint that has been employed to further reduce the number of incorrect matches is by analysing previously found disparities [CapPic00b]. During matching, numerous edge pixels can often be located from the same real world object. In this case, if two or more likely candidates are

found, then priority is given to the one that most closely matches the previous disparity that was correctly located. Figure (4.5) shows an example where 'A' indicates a positive match between a pair of stereo images. Arrows 'B' & 'C' represent possible matches found while searching for the next correspondence. By considering the previous match ('A'), it can be assumed that 'B' is incorrect and that 'C' is valid and belongs to the same real world object as 'A'. Figure (4.6a) portrays a depth map generated using this technique. This simple matching constraint was used instead of the PMF algorithm (section (4.2.3)) since it provided similar results in terms of quality of depth map, but at superior speeds, as shown later in section (5.1.3).

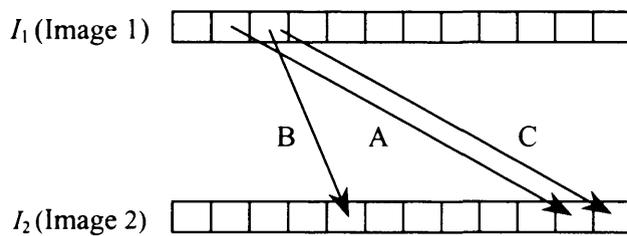


Figure (4.5) – Stereo matching.

During this process the third camera can be used to verify the integrity of any possible match by simply searching for similar edge properties in the location that corresponds to the current pixel disparity. If the match is not corroborated, it is assumed to be erroneous, and the search is resumed until either a match is located, or the maximum search range is reached and the pixel is deemed invalid, possibly due to an occlusion. An example of an edge depth map generated in this way is shown in figure (4.6b). Comparing figures (4.6a) & (4.6b) it is apparent that using a third camera for verification of matches greatly improves the accuracy over standard two camera techniques.

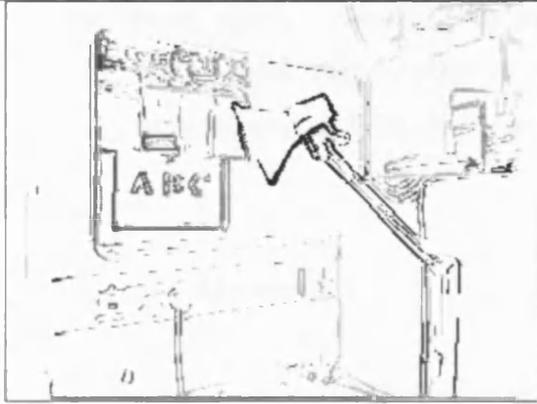


Figure (4.6a) – Standard edge depth map generated from figures (4.2a) & (4.2b), using a disparity range of 14-30 pixels.

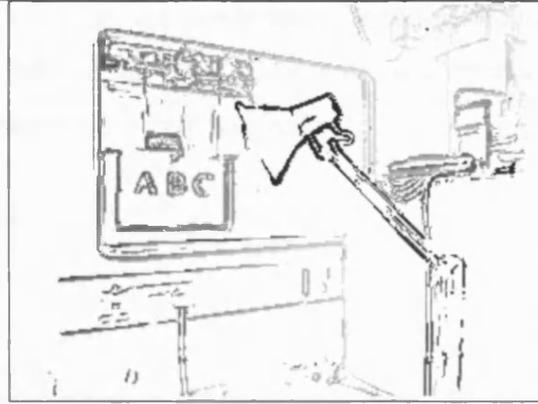


Figure (4.6b) – Standard edge depth map generated from the three stereo images of figures (4.2a) & (4.2b), and figure (4.4a), with a disparity range of 14-30 pixels.

4.3.4. Fast stereo edge depth maps

Whilst studying methods for increasing the frame rate of stereo techniques, an original technique was developed that generates a depth map consisting of edges that vary in width depending on the distance of objects in the scene. The closer an object to the camera, the wider the resulting edge. The limitation of this technique is that the stereo images used must consist of only small disparities. However, since the stereo algorithm consists of only binary operations, the depth maps can be processed at rates that are over an order of magnitude faster than most standard techniques. Trials on images with dimensions of 320x240 pixels demonstrated processing times of 8ms and lower to generate the edge depth maps.

It should be noted that the investigation into stereo techniques was not directly intended to invent new methods of stereo processing, but rather to modify existing methods in the hope of creating a more advantageous form of presentation. The technique described below was a new, albeit simple, procedure for generating stereo depth maps, but it does not compare with the generally accepted interpretation of a depth map. The resulting resolution is quite low, and would generally be considered too inferior for use in many applications (with the exception of tasks that require

extreme frame rates). However, as an input to the optophonic scene-to-sound mapping, with which the purpose is to provide a blind user with information necessary for enhanced mobility, it is considered more than adequate.

4.3.4.1. Algorithm

The proposed method [CapPic00a], founded upon a time-varying edge detector [HayJai83, Pic89], generates a very fast, low-resolution edge depth map. The technique was based in part, on a stereo-optical edge detector that was proposed by Avery Johnson [Joh63] in 1963, who believed it would be possible to use a small mosaic of photosensitive elements to locate nearby objects (contrast boundaries). The system, whilst moving with a translational motion, would detect only the closest objects in the scene. As the device containing the array of photosensitive cells moved, nearby objects would be observed as moving with a greater velocity than distant objects. Consequently, the light from the closest objects would cross the greatest number of the device's array elements, whereas distant objects would not cross any cells in the mosaic, hence not registering.

Assume that a pair of stereo images, labelled Frame1 and Frame2, both have a size in the region of 320x240 pixels or less. A low pass filter is passed over the two images, followed by a large Prewitt or Sobel edge operator. While performing the edge detection, all edge pixels above a minimum threshold value are set to 1; all others are set to zero. Finally the depth map is generated from a series of 1's and 0's. This is achieved by setting each pixel in the output image to one if the equivalent pixels in Frame1 and Frame2 are equal to zero and one, respectively. Otherwise, the output is set to zero. In summary:

**If ($S(\text{frame1}) > \text{min_threshold}$) then $S(\text{frame1}) = 1$
Else $S(\text{frame1}) = 0$**

**If ($S(\text{frame2}) > \text{min_threshold}$) then $S(\text{frame2}) = 1$
Else $S(\text{frame2}) = 0$**

**If ($S(\text{frame2}) == 1$ and $S(\text{frame1}) == 0$) then Result = 1
Else Result = 0**

Where S represents a 5x5 (or larger for thicker edges) edge operator, or any other suitable edge detector that is capable of producing relatively 'thick' edges. In the resulting image depth is (generally) proportional to the thickness of the edges, where a one signifies an edge.

The concept behind this technique is that a distant object generates a line that lies in approximately the same position in both frames. Therefore, if the two images were overlaid, a large or complete overlap of the two lines would be observed. The procedure would either reveal a set of zeroes, or a very thin line along the region where the two edges did not overlap. This effect reveals wider edges for objects nearer to the camera(s), since the greater the disparity between the edges, the less they overlap.

This procedure makes use of frames taken from a two-camera system, or alternatively, two consecutive frames captured from a monocular device that has undergone some form of translational motion.

4.3.4.2. Limitations and possible remedies

There are a number of restrictions and considerations to be taken into account when using this algorithm, but for use with a mobility aid, these are minor in comparison to its advantages, such as the speed of computation. Nonetheless, these limitations still exist.

An obvious drawback in the resulting depth maps is the loss of edges lying parallel to the epipolar line. However, this is a failing of many stereo depth maps, and is similar in nature to the aperture problem of cameras. For example, consider a scene with a long horizontal line that more than fills the view of an observing camera. Without texture in the scene it would be hard to identify any horizontal motion of the line. With the technique presented, if it encounters an edge parallel to the epipolar line, then it cannot determine the line's disparity, so the line will not be shown in the final image. This can be considered an advantage, since many stereo algorithms produce a large number of errors whilst attempting to find correspondences lying upon an edge that is parallel to the epipolar line. This form of error can be observed with the 'ABC' sign in figures (4.7c) & (4.7e) on page 88 that represent a modified version of the technique described here and the standard stereo edge technique previously illustrated. In figure (4.7d) the top and bottom edges of the sign are almost parallel to the epipolar line, and so are barely visible. In figure (4.7e) these same edges are portrayed as dark lines that incorrectly indicate a close proximity to the observing cameras.

Another problem with this and other stereo systems arises from repetitive patterns in the scenery. If two or more parallel lines are encountered in close proximity, this technique usually removes a line, or part of a line (depending on the disparity at that location), whilst displaying the rest. This only happens when the disparity is great enough to cause adjacent lines in the stereo image pair to overlap.

Not so much of a limitation, but a consideration, are the maximum recognisable disparities that this procedure exhibits. The maximum disparity correctly registered by this technique, in terms of pixels, is equivalent to the width of the edges in the image after applying the edge operator. (Note – any disparities greater than the edge widths, i.e. from objects lying very close to the cameras, will have the same disparity as indicated by the line widths themselves). Consequently, the stereo cameras are restricted to being positioned very close together so the maximum stereo disparity encountered is kept small. In the case of a single camera for a monocular system, scenes have to be sampled at high frame rates, taking consecutive frames as the stereo pair of images. This limits the maximum possible movement of any object in the

scene between frames. Fortunately, with the speed of operation of this algorithm, computation at a high frame rate is acceptable.

The use of a low pass filter before applying the edge detector to the images can lessen the effect of the above restriction. This is achieved by a widening of the edges in the images, as well as aiding in the removal of excess texture. Alternatively, advantages can be made by reducing the size of the initial images (prior to edge detection), for instance from 320x240 to 160x120 pixels, by taking the mean value of every 2x2 image block. This helps by performing a very effective low pass filter, and by further reducing the necessary computation time for the remainder of the procedure. In addition, the resulting depth map is visually more acceptable due to the ratio of the maximum line width to image size. As the image is reduced in size, the thickest lines (those depicting the greatest disparities) are represented more clearly. For example, a line of 10 pixels in width would be more noticeable in an image of size 160x120 pixels than in one of 320x240 pixels. This assumes that the image lines demonstrate a near constant thickness before and after the process of image reduction, which is generally the case, since the process of mean value reduction effectively spreads the majority of edges. Whether the image is 320x240 or 160x120 pixels in size, the maximum edge thickness obtained depends on the size of the edge operator (a 5x5 operator forming an edge with a maximum width of about 5 pixels), or on any additional methods of edge thickening that are employed.

An additional limitation that only arises when using a monocular camera system is the need to restrict camera motion to purely translational movement. This is a common failing of monocular camera systems, which can often result in a reversed depth map, or worse. If the camera undergoes a rotation along the horizontal (such as the holder of the camera turning round), then distant objects will appear to have a greater disparity than those closer at hand.

4.3.4.3. Results

The maximum line width generated by an enlarged (5x5) Prewitt operator in the example images was six pixels. This means that the depth map in figure (4.7c) has six

(seven if the lack of a line is included – white) possible ranges or depths that correspond to the six line widths, which for the purposes of developing a blind mobility aid, is more than adequate.

Most stereo edge depth maps consist of lines of varying intensity, whereby the intensity is directly proportional to the distance of a real world object, instead of varying edge thickness. Consequently, it was felt necessary to include some sort of comparison. To accomplish this a simple automated system has been incorporated that rapidly converts the line width into a corresponding intensity scale (figure (4.7d)). Since there are seven different line widths (0, 1, 2, 3, 4, 5 & 6 pixels) in the depth map, the corresponding pixel intensities range from 0 to 255: 0 (foreground), 42, 85, 127 (middle distance), 170, 212, 255 (background). Where an intensity of zero corresponds to a black line six pixels in diameter, and an intensity of 255 represents a white pixel, or the lack of a line.

As a guide, the depth map obtained from a simple stereo edge detector has been included (figure (4.7e)). This detector used the Sobel (or Prewitt) edge operator to obtain the local edge magnitudes and orientations from the stereo image pair and then performs matching by searching for candidates that have near identical edge properties within a given disparity range, as explained in section (4.3.3). A candidate match for this stereo procedure between the two images is one in which the difference in edge orientation is less than 30 degrees, and the magnitudes vary by no more than 30 (from an intensity range of 0-255).



Figure (4.7a) – Left frame of a stereo pair.



Figure (4.7b) – Right frame of a stereo pair.

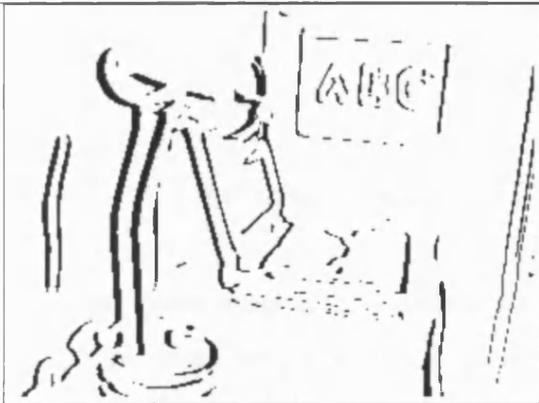


Figure (4.7c) – Rapid stereo depth map. Depth portrayed by relative width of edges in scene. Main errors caused by edges lying parallel to camera baseline.



Figure (4.7d) – Rapid depth map converted into intensity edge depth map, allowing comparison with edge depth map of figure (4.7e).

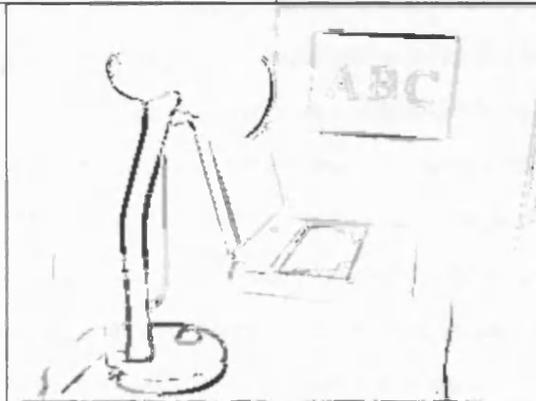


Figure (4.7e) – Simple stereo edge image for visual comparison with figure (4.7d). Errors occur with edges lying parallel to camera baseline.

The depth map resulting from this technique is quite suitable for conversion into sound for a number of reasons. Objects close to the observing camera(s) are represented with thicker edges, and after conversion into sound, they generate a tone for a longer time period. If the intensity modification is incorporated as well, then the closest objects (those of greatest significance for immediate mobility) will generate louder tones and more effectively capture the user's attention. Distant objects, which are represented by light thin edges and are generally of less significance, will be portrayed with short quiet tones and will not distract the user from more important features.

4.3.5. Stereo cartooning

An edge depth map is one solution to the task of improving the output of an optophonic blind aid by placing emphasis on regions of high interest, such as object boundaries that lie close to the observing cameras. Realistically an edge image is not the ideal solution. It can often be very difficult to interpret even relatively simple scenes when portrayed in the form of an edge depth map (figures (4.6a) & (4.6b) and (4.7c), (4.7d) & (4.7e)). For example, objects can become hard to recognise purely from their outlines alone, especially if parts of that outline are missing. Pearson & Robinson [PeaRob85], and more recently Manoranjan [Man98], encountered a similar problem whilst attempting deaf sign communication over the phone network via real-time low-resolution video. Due to necessary bandwidth reductions, normal low-resolution video images occupied too much space in terms of data, and could not be displayed at fast enough frame rates for accurate perception of hand signing. While edge images could be compressed quite adequately to allow for in excess of 8 frames per second, which was sufficient to capture movement or signing. The loss of image detail due to edge detection made the accurate perception of signing very laborious and generally near impossible.

The proposed solution used two-tone cartoon images (figures (4.8a) & (4.8e)) that contained sufficient detail to allow for accurate interpretation of the hand signs, whilst suppressing enough unwanted scene information to relay the video at a comfortable frame rate for perception of movement.

The next method presented was not intended as a new stereo technique. It was designed for incorporation and use with any current stereo algorithm, to enhance depth maps ready for conversion into sound via the optophonic mapping. The process achieves this by attempting to restore lost texture (such as text) and shading back into the final disparity image, by combining an edge depth map with the two-tone cartoon technique. The result is a cartoon-like edge depth map [CapPic00c, CapPic00d] that greatly reduces the emphasis on unnecessary features within the scene, whilst retaining object and region structure.

The cartoon image technique simply applied a threshold to an image. Any pixel intensities darker than the threshold were replaced with black pixels, all others were set to white. This, when combined with an edge detection routine, resulted in a two-tone cartoon filled image.

The threshold is determined using a method proposed by Manoranjan, whereby the absolute threshold is determined, via histogram, using a fraction of the cumulative sum of pixel intensities in the image scene. Although this technique is not ideal, it does provide an adequate solution for most real world scenes. (Appendix 3 – section (A3.3.3), shows how the histogram thresholding can be optimised for faster processing).

The cartooning method was adapted to stereo images, so that edges and some surfaces would be displayed with a pixel intensity that was approximately proportional to the distance from the cameras (figures (4.8b) & (4.8f)). First, the original cartoon technique is applied to the left stereo image, producing an image with edges and surfaces (figures (4.8a) & (4.8e)). It is advantageous to set a third tone (other than black or white) in the cartoon that corresponds to object edges, used in preventing region filling beyond object boundaries. Secondly, an edge operator is applied to the right stereo image. A match is then found between an edge pixel in the left and right

stereo images, and an estimate made of the depth from the camera. The equivalent pixel in the edge depth map is then encoded with a brightness corresponding to its depth. Using the cartoon image (the left stereo image) any surface that has been coded black can be re-coded using the same pixel intensity as the edge that has just been found, since it is assumed that the edge and the surface are of the same depth. This ceases when another match is found between the stereo images, which is assumed to be the opposite end of the surface.

For example:

Assume that the cartoon and depth map images are represented by the two-dimensional arrays `cartoon_image[vertical][horizontal]` and `depth_image[vertical][horizontal]`, respectively, and that black, when considering the depth image, indicates the lack of any currently discovered disparity. Whilst in terms of the cartoon image, a black pixel represents part of a filled area. A third pixel tone being used in the cartoon image to correspond to object edges, which is used to prevent region filling beyond object boundaries.

```
while (cartoon_image[vert][l_horz]==black & depth_image[vert][c_horz]==black)  
    depth_image[vert][c_horz]=intens  
    l_horz=l_horz-1  
    c_horz=c_horz-1
```

In the above routine, pixel filling ceases in the depth image when either an edge boundary is encountered, or the current pixel in the cartoon image is unfilled. This generally indicates the start or end of an object, and a possible change in depth.

The depth map that this technique generates represents the scene in the form of a cartoon, with objects being displayed in various shades of grey that correspond to their distance from the observing cameras (figures (4.8b) & (4.8f)). This aids in visually identifying scene objects, by defining regions of similar shading within the image. Similarly, the sound output generated from the resulting cartoon depth map produces a fuller sound for close 'filled' objects, than that from the equivalent edge depth map. This depth map is a cross between an edge depth map and an intensity depth map, having the advantages of presenting text, whilst retaining areas of filled regions that effectively portray objects.

The stereo cartoon system described is simple enough to work at speeds in excess of 15 frames per second, with adaptive cartoon thresholding. Once generated, this cartoon depth map is converted into sound via a standard optophonic mapping, for interpretation by the listener/user. This method successfully reduces the quantity of relatively nonessential features within the scene, such as texture, whilst providing the user with depth and object information via shading.

This method for presenting images makes use of any stereo algorithm and so accurate depth information could be obtained using more up-to-date stereo techniques. However, with technology at its current level, even a high specification PC would be unable to process the images, generate the depth maps, and perform the scene-to-sound mapping at sufficiently high speeds using a more complex stereo algorithm. Using the stereo edge algorithm (section 4.3.3) with this form of presentation (cartooning) and the inclusion of the scene-to-sound mapping, speeds of 12-15 frames per second are obtainable (image dimensions of 320x240).

It should be noted that the cartooning algorithm has been unable to correctly classify the foreground (the table at the very front/bottom of the image) in figure (4.8f), which rather than being portrayed as close to the camera (which for reasons of visual clarity white pixels represent close objects) it has been represented as distant (black pixels). This has occurred due to the lack of detected stereo edges in this region of the scene. If insufficient object boundaries are detected then the stereo cartooning technique does not perform region filling since in the majority of cases this implies that no object exists in that region. However, this is not so in the example shown in figure (4.8f), which is caused by an object being very close to the observing cameras (thus filling a large region of the captured image with very few edge boundaries). Fortunately in these circumstances the user would be so close to the object that they would most likely already be in contact with it, after previously being alerted to its presence whilst at a distance.

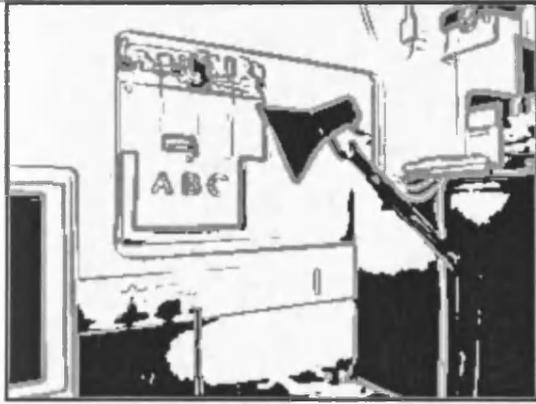


Figure (4.8a) – Two-tone cartoon edge image generated from figure (4.2a) on page 72. Three tones have been used for clarity. Filled regions are black (intensity 0), and Sobel edges are grey (128). All other regions are white (intensity 255).



Figure (4.8b) – Cartoon-edge depth map generated from edge depth map, figure (4.6a) on page 82, and the cartoon image, figure (4.8a). Closest objects are displayed in white for reasons of clarity.

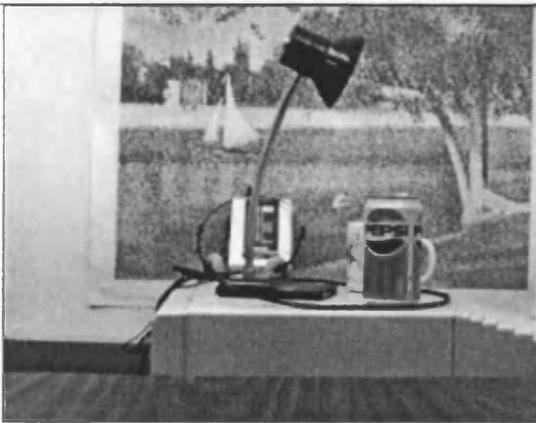


Figure (4.8c) – Left stereo image of a simple scene.

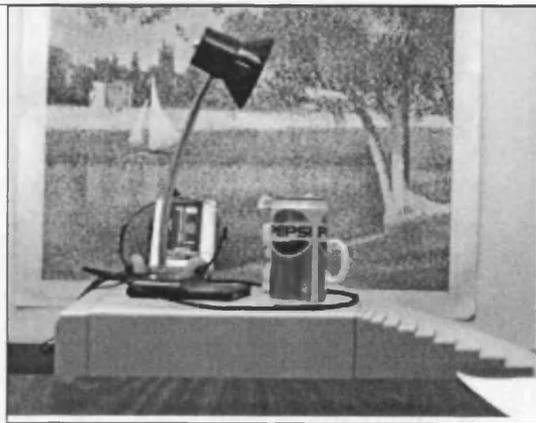


Figure (4.8d) – Right stereo image for the scene shown in figure (4.8c).

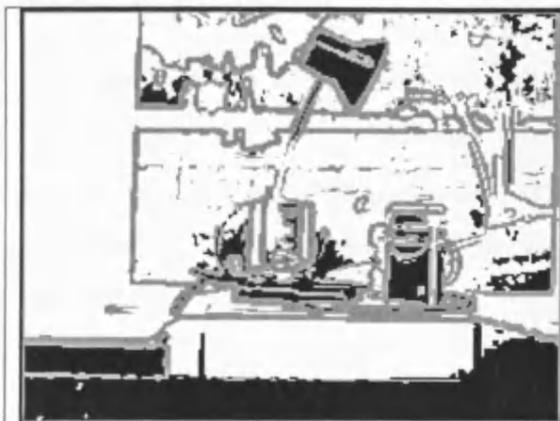


Figure (4.8e) – Another example of a cartoon image generated from figure (4.8c).

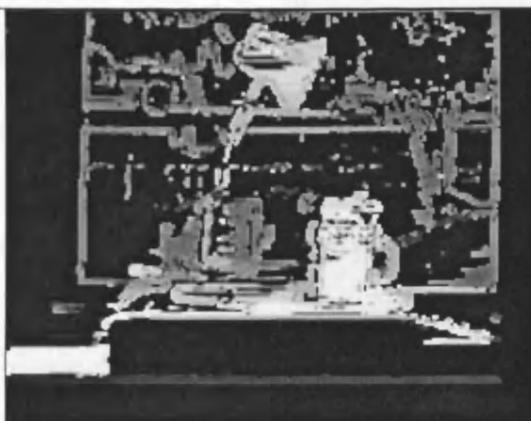


Figure (4.8f) – Cartoon depth map generated from figures (4.8c), (4.8d), & (4.8e). Closest objects are displayed in white for reasons of clarity.

4.4. Summary

Before considering the use of a stereo algorithm to generate an image ready for conversion to sound, it was necessary to determine what image or sound properties should be sacrificed to make way for the depth information. If the relative pixel intensity in the depth map (and consequently the amplitude of the generated sound) were related to distance, such that a distant object would produce a much fainter sound than an object closer at hand, then this would not allow this parameter to convey textural detail. This limitation (the lack of textural detail) was accepted for the more desirable qualities of a depth map.

The first stereo techniques examined were the standard intensity based depth maps. Using any one of these procedures it is possible with only two adjacent image frames of a scene, with a known epipolar line, to obtain a good graphical representation, whereby pixel intensity is proportional to object distance. However, it was shown that this form of depth map has a limitation that results from its quality. The greater the quality, the greater the computation time, often in excess of several minutes on a standard computer for a high-resolution image.

Consequently, the investigation was directed towards the possible use of the faster stereo edge depth map. Although often visually inferior due to their restriction to regions of rapidly changing contrast, or edge boundaries, edge depth maps have the advantage that they can be processed at fairly high frame rates (multiple frames per second).

Whilst studying methods of increasing stereo frame rates, an original technique was developed. The depth map generated by this new method was shown to consist of edges varying in width depending on the distance of objects in the scene. The closer an object to the camera, the wider the resulting edge. This stereo algorithm consisted of only binary operations, so could be processed at rates of over an order of magnitude faster than most standard techniques.

Although a great deal of progress was made with stereo edge processing, the resulting images were generally considered to be inadequate for easy recognition of the scene. With conversion to sound the problem would only be worsened. To contend with this problem the work of Pearson & Robinson [PeaRob85] was considered, bringing about a new method for modifying the appearance of stereo depth maps by utilising a cartoon image. This method attempted to replace the lost regions of texture in the scene by shading areas with a pixel intensity corresponding to the relative distance from the camera. The final depth map is a cartoon that is a combination of an edge depth map and an intensity depth map. Not only does the method provide a visually acceptable result, but also the computation time is comparable to that of a standard stereo edge technique, processing multiple frames every second.

5. Optophonic performance – Stereo versus monocular vision

Due to the difficulties encountered whilst evaluating the effectiveness of a stereo depth map, which may involve comparisons with a hand-generated stereo image, an automated method was derived that provided accurate guidelines to the most suitable depth algorithm in terms of speed and efficiency. This method for evaluating the effectiveness of a stereo algorithm was later found to be similar to a technique called the False-Positive-Fraction. Using this technique it was possible to decide on the most suitable stereo algorithm for use with the optophonic mapping.

A series of experiments were devised that utilise a number of volunteers to gauge the effectiveness of the techniques outlined in previous sections compared to the standard optophone mapping. The tests, partially based upon a system of examination labelled DeLIA (Detection, Location, Identification, & Avoidance) that identifies key stages in the process of detecting the presence of possible obstacles, were designed so that an indication could be provided as to whether the listener more readily perceived a particular form of image. Data gathered would then indicate the usefulness of the proposed techniques and provide a deeper understanding of the requirements necessary to create a suitable mobility aid for the blind.

Conclusions drawn from the results give an indication as to what parts of an image the blind user could possibly manage without, and the regions that should be emphasised for easier recognition.

5.1. Evaluation of depth maps

There are few techniques used for evaluating the effectiveness of a stereo depth map, mainly because they are difficult to assess. A method of evaluation may work well when assessing one particular stereo algorithm, however, it may fail to successfully evaluate another. Intensity depth maps can be judged in a different way to an edge depth map, since for the latter an indication of accuracy is provided by the quantity of

complete edges. Any break-up of the lines within the depth map signifies probable errors in matching. On the other hand an intensity depth map should generally consist of large regions of similar intensities, thus a speckled depth map may imply the presence of inaccuracies.

One method of evaluation that seems to be more universal, albeit somewhat less analytical, is that of visual inspection. Another factor considered in the determination of a stereo algorithm's performance is its speed. An ideal algorithm would produce a high quality depth map as quickly as possible. Alternative methods of evaluating stereo algorithms do exist, but these tend to be time consuming, involving a great deal of manual work, such as the determination of correspondences by hand [Fua93]. These selected matches are compared with results obtained from the stereo algorithm under trial and used to gauge its performance. The accuracy and reliability of this form of evaluation is proportional to the sample of handpicked matches taken.

5.1.1. False Positive Fraction

A faster, more suitable and automated method for determining the best procedure for detecting and generating an edge depth map was developed during the research, and was later found to be similar to a technique called the False Positive Fraction (FPF) [MetPan99, SkuConGor99]. FPF corresponds, in this case, to the fraction of positive matches made by the stereo algorithm that are actually falsely classified.

The method used to evaluate the quality of a stereo edge algorithm requires a pair of stereo images in which the largest disparity in pixels is known, or can be determined. For example, the greatest disparity of any object clearly visible in both figures (5.1a) & (5.1b) (page 129), belongs to the head of the lamp, and is 30 pixels. Similarly, the smallest disparity between the two figures was found to be 15 pixels. Next the edge depth map is generated using the chosen algorithm. For the evaluation technique to work, the disparity range used must be from the lowest disparity (in this case, 15 pixels), or lower, to one that exceeds the maximum disparity found between the stereo image pair (for example - 40 pixels). The algorithm is then applied so that the final edge depth map consists of edges with pixel intensities varying from black through to

light grey. White being used to represent areas where no edge matches are located. Finally a histogram of pixel intensities is taken from the edge depth map.

From the histogram, all pixel intensities that correspond to matches that exceed a disparity of 30 pixels are known to be incorrect (the FPF). Also, the total number of pixels corresponding to edges in the depth map is known. From these a ratio can be obtained of the total minus the incorrect matches to the total matches alone over the given range of disparities. During this procedure a number of assumptions and observations were made. For example, as the maximum disparity increases, the number of false matches generally increases. Furthermore, by setting the lowest allowable disparity below any that exist between the stereo pair more false matches may have been encountered. This quantity was generally very low since the stereo techniques tested were designed so that in situations where several equally likely matches were found they would either be ignored or the one that provided the greatest disparity chosen. The important factor to remember is that the FPF is only used as a guide to the ideal technique. Consequently, as long as all stereo algorithms under evaluation are tested using the same properties, a good indication of the algorithm's performance is obtained, with a perfect algorithm receiving a ratio of 1.

For example, consider the simple stereo algorithm (section 4.3.3), which uses previous disparities as a guide as well as removing previously matched pixels from future searches. This algorithm, labelled A1R1C0S0T0, when tested both visually (figure (5.1w) – page 131) and via the FPF method (table (5.1)) provided some of the best results in the shortest computation time. From the FPF results it is shown that only 4% of the matches are known to be incorrect, giving a ratio of $(12348 - 496)/12348 = 0.96$, or 96% for the 'Assumed Correct Matches' value. This method has been applied to a number of images with a variety of content, texture, and illumination, and the FPF results obtained were accurate to within a few percent of the figure quoted. Two sets of sample data can be seen in tables (5.2) & (5.3) on pages 104 & 105, which correspond to the images shown in figures (5.1c)-(5.1z) & (5.2c)-(5.2z) (pages 129-136), respectively.

Disparity Range	Number of Pixels	Pixel Type
0	64452	Blank pixels
15-30	11852	Other matches
31-40	496	False matches
15-40	12348	Total matches

Table (5.1) – The results of applying the described method for evaluating stereo edge depth maps to the procedure labelled A1R1C0S0T0, shown in figure (5.1w).

Finding the number of edge pixels in the stereo edge images provides a further refinement to the method of assessing performance. For figures (5.1a) & (5.1b) (page 129) the number of edge pixels found after applying an edge operator was approximately 18200 pixels. The edge depth map (A1R1C0S0T0) contained 11852 pixels (table (5.1)) that corresponded to matched edges (assumed correct), which is only 65% of the original edge pixels. This is explained by the fact that the extreme left and right borders of both images cannot be used for stereo matching due to lack of image overlap. Excluding these borders from the count, the percentage of edges that appear in the edge depth map is over 70%. This value ('Matched Edge Pixels') provides an estimate as to the percentage of edge pixels that are correctly matched during the stereo process, whereas the percentage of matches assumed to be correct gives a fairly accurate indication to the quantity of correct matches in the depth map.

5.1.2. Stereo algorithms

Tables (5.2) & (5.3) (pages 104 & 105) illustrates results of using the FPF as a gauge for assessing the accuracy of stereo edge algorithms. The five techniques that were tested individually and in combination with each other are:

(A) - Advanced Search – Using previously found disparities as a guide for future matching.

(C) - Contrast Matching – As well as matching edge magnitude and orientation, use the change in contrast.

(R) - Remove Matched Pixels – Once a likely match has been made, remove the matched edge pixels from future searches.

(S) - Select Best Fit – Not used in conjunction with (A). Rather than taking the first likely match, or using previous disparities as a guide, find all candidate matches within the given disparity range and select the best fit with respect to edge orientation and magnitude.

(T) - Twin Search – Search from image A to B, then verify matches with those found by searching from B to A. If not verified in both searches then remove the match.

The basic stereo algorithm used the edge orientation and magnitude obtained from a Sobel edge operator and is represented in the tables as A0R0C0S0T0, where the zeros represent the lack of use of a particular option.

From the tables, the False Positive value corresponds to the quantity of known false matches (lying outside the maximum disparity that exists between the stereo images) made by the stereo algorithm.

The Assumed Correct Matches value corresponds to the number of pixels that have been matched within the known correct disparity range as a percentage of the total matches made.

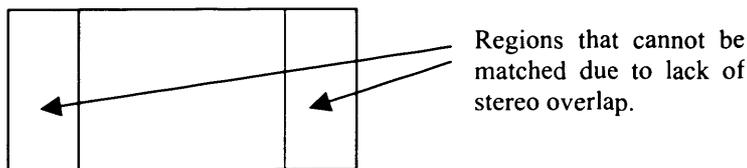


Figure (5.3) – Diagram showing the regions that cannot be matched due to lack of stereo overlap.

The value for Matched Edge Pixels relates to the number of pixels that are assumed correct as a percentage of the number of edge pixels in the original images. A more accurate (and favourable) value would be provided by using the number of edges in the modified edge image, which takes into account the areas of the images that cannot

be matched due to the stereo disparity (figure (5.3)). However, this change would not alter the ranking of the stereo algorithms in terms of performance.

5.1.3. Stereo performance

The described algorithms were evaluated with numerous images consisting of various objects, shading and lighting, texture, and other factors. In each case the order of performance (indicated by the algorithm order in tables (5.2) & (5.3)) varied little with only marginal variation in percentage scores, with the overall best algorithm combination being A1R1C0S0T0 (advanced search and matched pixel removal only). This method produces reasonably fast and accurate edge depth maps.

The speeds shown for the algorithms in tables (5.2) & (5.3) correspond to 320x240 pixels images and were obtained by performing the testing on a 366MHz AMD processor PC. The final set-up used for generating the optophonic signal captured via the Triclops camera system consisted of a P3 450MHz PC, giving a speed increase of 50-75%.

Although the results for the two-way search do not look promising in terms of the number of edge pixels matched in relation to the total that it is possible to match, nearly all correspondences found are correct. In some circumstances this situation is advantageous – finding a small percentage of the pixels, but having the almost certain guarantee that they are all correct. However, for use with a blind mobility aid the requirement is for fast depth maps that portray large numbers of pixels with a reasonable amount of accuracy, as provided by the algorithms that use previous disparities as a search guide. This is analogous to a stereo technique that demonstrates a high percentage for the ‘Assumed Correct Matches’ (matched pixels assumed correct) and ‘Matched Edge Pixels’ (all edge pixels correctly matched), as well as a high frame rate. As an example tables (5.2) & (5.3) have been sorted under the column that portrays the percentage of edge pixels correctly matched (labelled ‘% - Matched Edge Pixels’).

The method labelled 'S', which searches for all possible candidates within the given disparity range and then selects the most likely, was found to be rather slow (3 or 4 frames per second) and had extreme difficulties interpreting disparities for lines lying parallel to the epipolar line. The lack of speed is understandable when considering that the algorithm scans every pixel within the disparity range, rather than searching until the first likely candidate is located. Similarly, the rather drastic errors encountered with lines close to the epipolar direction are caused by the similarity of every pixel within the line. Without the use of previous disparities as a guide it becomes difficult to determine the correct match.

Incorporating contrast checking (labelled 'C') was found to speed up the stereo process by about 10-15% merely by using the direction of contrast change of edge pixels as an initial guide. Using this technique it was observed that the percentage of False Positives decreased by a fraction of a percent, however, the converse was observed to occur with the percentage of total edge pixels correctly matched. The reason for this appears to be that during an occlusion the contrast change at the edge of the foremost object can alter due to the differing background. Thus, the edge pixels may not pass the contrast check and so will not be matched.

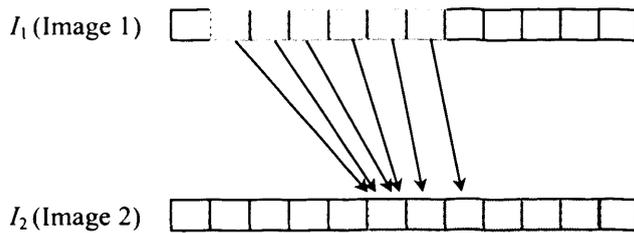


Figure (5.4) – Stereo matching (indicated by arrows) without the removal of previously matched edge pixels (edge pixels marked in grey).

The removal of successfully matched pixels (labelled 'R') improves the quality of the depth map by preventing some incorrect matching of future edge pixels, especially those along an edge parallel to the epipolar line. Without this removal, consecutive edge pixels would be matched with the first pixel of the edge in the second image (figure (5.4)). This diagram represents the matching process for lines parallel to the

epipolar line when previously matched pixels are not removed. Assuming a minimum allowed disparity of one pixel, each edge pixel would be matched as indicated by the arrows in the diagram.

The most useful technique that was tested, labelled 'A' (section (4.3.3)), used previous disparities to find possible alternative matches. This was used as a faster alternative to the PMF algorithm (section (4.2.3)), using the concept that edge pixels belonging to the same region or object would all have similar disparities. The algorithm, which demonstrated some of the best results in terms of speed, accuracy, and visual performance, avoided a number of the limitations of the PMF technique, such as mismatching caused by rapid changes in depth within the real world scene. For comparison, the PMF algorithm was tested and found to provide little or no improvement over the more simplistic approach.

The PMF can be seen to fail in some areas with the A and B shown in figure (5.5b) – page 137, (mainly on the left side of the B) due to the severe changes in depth. However, in the lamp example (figure (5.5a), page 137) the opposite is observed with a marginal improvement in quality (albeit extremely slight – a fraction of a percent when assessed using False Positives as shown in table (5.2)). Another drawback with this technique is the reduction in frame rate (a decrease in speed in excess of 15% as shown in tables (5.2) & (5.3)). Consequently, due to the limited improvement in quality, difficulties in perceiving rapid changes in depth, and the loss of frame rate, this technique was discarded in favour of the one described for checking previous disparities.

Table (5.2) - Lamp-Office-ABC Image Set - Stereo Algorithm Test

Combination Number	Algorithm Combination	Disparity Range →					0	15-30	31-40	15-40	%	%	Time	Speed
		Algorithm Characteristics					Blank Pixels	Other Matches	Known False Matches – False Positive	Total Matches	Assumed Correct Matches	Matched Edge Pixels	Seconds to process 200 images	Frames per second (320x240 images)
		(A) - Advanced Search - Use previous disparities as a guide	(R) - Remove matched pixels from search	(C) - Contrast matching	(S) - Select best fit from all candidate matches	(T) - Twin search - Search from A to B, then B to A								
Left Edge Image	Left edge image consists of 58598 blank pixels and 18202 edge pixels →					58598	18202							
Modified Left Edge Image	Modified edge image (accounting for stereo overlap where pixels cannot be matched) consists of 54868 blank pixels and 17132 edge pixels →					54868	17132							
Image Pixels	Total number of pixels in image is 320x240 →					76800								
2	A0R0C0S0T1	0	0	0	0	1	73194	3175	431	3606	88.05%	17.44%	54	3.70
6	A0R0C1S0T1	0	0	1	0	1	72956	3372	472	3844	87.72%	18.53%	49	4.08
18	A1R0C0S0T1	1	0	0	0	1	72708	3707	385	4092	90.59%	20.37%	57	3.51
20	A1R0C1S0T1	1	0	1	0	1	72522	3858	420	4278	90.18%	21.20%	50	4.00
11	A0R1C0S1T0	0	1	0	1	0	66700	5976	4124	10100	59.17%	32.83%	40	5.00
15	A0R1C1S1T0	0	1	1	1	0	66799	6009	3992	10001	60.08%	33.01%	37	5.41
3	A0R0C0S1T0	0	0	0	1	0	65733	6488	4579	11067	58.62%	35.64%	44	4.55
7	A0R0C1S1T0	0	0	1	1	0	65847	6527	4426	10953	59.59%	35.86%	40	5.00
16	A0R1C1S1T1	0	1	1	1	1	68828	7340	632	7972	92.07%	40.33%	91	2.20
12	A0R1C0S1T1	0	1	0	1	1	68685	7458	657	8115	91.90%	40.97%	108	1.85
8	A0R0C1S1T1	0	0	1	1	1	67478	8412	910	9322	90.24%	46.21%	102	1.96
4	A0R0C0S1T1	0	0	0	1	1	67406	8463	931	9394	90.09%	46.49%	121	1.65
22	A1R1C0S0T1	1	1	0	0	1	67535	8678	587	9265	93.66%	47.68%	62	3.23
24	A1R1C1S0T1	1	1	1	0	1	67463	8763	574	9337	93.85%	48.14%	56	3.57
10	A0R1C0S0T1	0	1	0	0	1	67206	8898	696	9594	92.75%	48.88%	62	3.23
14	A0R1C1S0T1	0	1	1	0	1	66935	9156	709	9865	92.81%	50.30%	56	3.57
5	A0R0C1S0T0	0	0	1	0	0	67141	9411	248	9659	97.43%	51.70%	21	9.52
1	A0R0C0S0T0	0	0	0	0	0	67111	9463	226	9689	97.67%	51.99%	22	9.09
13	A0R1C1S0T0	0	1	1	0	0	66014	10303	483	10786	95.52%	56.60%	21	9.52
9	A0R1C0S0T0	0	1	0	0	0	65950	10369	481	10850	95.57%	56.97%	22	9.09
19	A1R0C1S0T0	1	0	1	0	0	65121	11396	283	11679	97.58%	62.61%	22	9.09
17	A1R0C0S0T0	1	0	0	0	0	65007	11550	243	11793	97.94%	63.45%	23	8.70
23	A1R1C1S0T0	1	1	1	0	0	64600	11667	533	12200	95.63%	64.10%	22	9.09
21	A1R1C0S0T0	1	1	0	0	0	64452	11852	496	12348	95.98%	65.11%	23	8.70
PMF	A1R1C0S0T0	1	1	0	0	0	64343	12011	446	12457	96.42%	65.99%	27	7.41

Table (5.3) - A150cm_l-B75cm_r Image Set - Stereo Algorithm Test

Combination Number	Algorithm Combination	Disparity Range →					0	11-36	37-46	11-46	%	%	Time	Speed
		Algorithm Characteristics					Blank Pixels	Other Matches	Known False Matches – False Positive	Total Matches	Assumed Correct Matches	Matched Edge Pixels	Seconds to process 200 images	Frames per second (320x240 images)
		(A) - Advanced Search - Use previous disparities as a guide	(R) - Remove matched pixels from search	(C) - Contrast matching	(S) - Select best fit from all candidate matches	(T) - Twin search - Search from A to B, then B to A								
Left Edge Image	Left edge image consists of 58276 blank pixels and 18524 edge pixels →					58276	18524							
Modified Left Edge Image	Modified edge image (accounting for stereo overlap where pixels cannot be matched) consists of 56016 blank pixels and 17424 edge pixels →					56016	17424							
Image Pixels	Total number of pixels in image is 320x240 →					76800								
6	A0R0C1S0T1	0	0	1	0	1	72929	2707	1164	3871	69.93%	14.61%	47	4.26
2	A0R0C0S0T1	0	0	0	0	1	73019	2975	806	3781	78.68%	16.06%	54	3.70
18	A1R0C0S0T1	1	0	0	0	1	72545	3056	1199	4255	71.82%	16.50%	56	3.57
20	A1R0C1S0T1	1	0	1	0	1	72391	3214	1195	4409	72.90%	17.35%	50	4.00
11	A0R1C0S1T0	0	1	0	1	0	67455	5238	4107	9345	56.05%	28.28%	45	4.44
15	A0R1C1S1T0	0	1	1	1	0	67518	5255	4027	9282	56.61%	28.37%	41	4.88
3	A0R0C0S1T0	0	0	0	1	0	65474	6186	5140	11326	54.62%	33.39%	52	3.85
7	A0R0C1S1T0	0	0	1	1	0	65534	6213	5053	11266	55.15%	33.54%	48	4.17
16	A0R1C1S1T1	0	1	1	1	1	69788	6426	586	7012	91.64%	34.69%	93	2.15
12	A0R1C0S1T1	0	1	0	1	1	69749	6467	584	7051	91.72%	34.91%	109	1.83
8	A0R0C1S1T1	0	0	1	1	1	68411	7407	982	8389	88.29%	39.99%	105	1.90
4	A0R0C0S1T1	0	0	0	1	1	68387	7430	983	8413	88.32%	40.11%	123	1.63
5	A0R0C1S0T0	0	0	1	0	0	68608	8104	88	8192	98.93%	43.75%	22	9.09
1	A0R0C0S0T0	0	0	0	0	0	68557	8152	91	8243	98.90%	44.01%	24	8.33
22	A1R1C0S0T1	1	1	0	0	1	67034	9410	356	9766	96.35%	50.80%	64	3.13
10	A0R1C0S0T1	0	1	0	0	1	66905	9441	454	9895	95.41%	50.97%	64	3.13
24	A1R1C1S0T1	1	1	1	0	1	66969	9456	375	9831	96.19%	51.05%	57	3.51
14	A0R1C1S0T1	0	1	1	0	1	66876	9489	435	9924	95.62%	51.23%	57	3.51
13	A0R1C1S0T0	0	1	1	0	0	66082	10574	144	10718	98.66%	57.08%	23	8.70
9	A0R1C0S0T0	0	1	0	0	0	65996	10656	148	10804	98.63%	57.53%	24	8.33
19	A1R0C1S0T0	1	0	1	0	0	65724	10965	111	11076	99.00%	59.19%	24	8.33
17	A1R0C0S0T0	1	0	0	0	0	65641	11034	125	11159	98.88%	59.57%	25	8.00
23	A1R1C1S0T0	1	1	1	0	0	64064	12564	172	12736	98.65%	67.83%	24	8.33
21	A1R1C0S0T0	1	1	0	0	0	63958	12660	182	12842	98.58%	68.34%	25	8.00
PMF	A1R1C0S0T0	1	1	0	0	0	65002	11612	186	11798	98.42%	62.69%	30	6.67

5.2. Optophonic testing

After determining the most suitable stereo edge algorithm in terms of both visual appearance and speed, as well as with the previously described method for evaluating the efficiency of stereo processes, the technique was combined with the cartoon depth map. This method, due to the quality of the results obtained, was used for further testing with volunteers.

To determine how to assess the various modifications to the optophonic system and to decide upon what form preliminary tests should take, the various aspects and thought processes that are required for general mobility were broken down into categories.

These stages include:

- **Detection** – The first stage involves recognising that there is an object in the immediate vicinity.
- **Location** – Once it is known there is a nearby obstacle or object, determine its location.
- **Identification** – Next it is important to determine what the object is.
- **Action/Avoidance** – After determining the identity of the object, action may need to be taken such as the avoidance of an obstruction.

The proposed method of evaluation was labelled DeLIA for the Detection, Location, Identification and Avoidance of obstacles, representative of the process with which a blind user might attempt to navigate through an unfamiliar environment using a mobility aid. Hence, a series of tests were created using volunteers in an effort to assess each of the stages involved in DeLIA.

5.2.1. Initial experiments

The first two tests described in this section are rather subjective and were carried out with 3 volunteers to gain insight into the optophone. The first was used to establish whether stereo sound was indeed useful and to get a feel for how accurately frequency could be used to judge vertical position. The second test was used to obtain an impression of what it is like to navigate with the optophone. Neither test is used to establish the overall effectiveness of the optophone, but helps to provide clues as to what further work could be undertaken.

5.2.1.1. Test1 – Frame rate

Optophonic devices are commonly set to a display time of one frame every 2 seconds, which provides the listener with enough time to perceive prominent features within a simple image scene. However, it was observed that similar results could be obtained with a slightly faster scan rate of 1 or 2 frames per second after a longer training period (in the region of twenty minutes). When using stereo depth maps with the optophonic mapping, faster display rates sometimes appeared to provide enhanced depth clues over the longer two second scan, probably due to changes in depth appearing to have a greater emphasis with the faster scan.

At an early stage in the research the frame rate was considered an important aspect. It was unknown whether the human brain could be trained to perceive a rapid frame rate, when played as sound, in a similar fashion to that of television. For example, beyond ten frames per second the human visual system starts to perceive fluid motion (the Gestalt effect). However, even using the stereo techniques and the other processing methods described, four frames per second appeared to be the maximum rate readily comprehensible without excessive training.

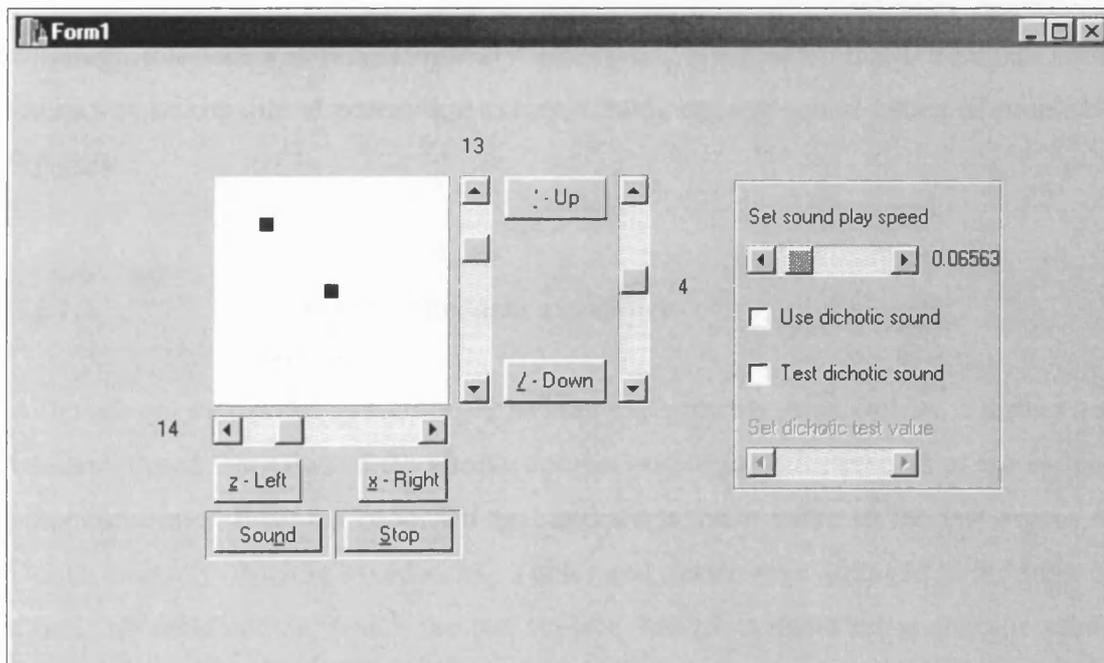


Figure (5.6) – Screenshot of the programme used to test a subject's ability to perceive the location of two moveable dots by sound alone at high frame rates.

Even after this apparent ceiling for intelligibility was found for real-time optophonic processing, it was believed important to study the effects of increasing frame rates for basic images. For this reason an optophone programme was written to play sound that corresponded to an image consisting of two small black dots (4x4 pixels in size within an image of 64x64 pixels) that could be manoeuvred around a white background (a screenshot of this programme is shown in figure (5.6)). The image scene was hidden from view, with the two dots being randomly positioned within the image. Listening to only the sound relating to the hidden image the subject was asked to move the second dot over the first using keyboard controls for the vertical and horizontal directions. Once the subject believed they had accomplished the task the image would be displayed on the screen.

With respect to the horizontal location even at speeds of 16 and 32 frames per second subjects were able to accurately align the dots. The vertical positioning on the other hand appeared to be slightly more random at these high frame rates. The only way the height could accurately be determined was by shifting the moveable dot between the extremes (high frequency to low frequency, top to bottom, respectively) and then gauging the difference.

Although this test was not completely successful, it did show that the human brain appears to be capable of perceiving extreme frame rates of sound (albeit of simplistic images).

5.2.1.2. Test2 – Obstacle avoidance

Although not monitored as rigorously as later experiments using DeLIA, a further test was performed that allowed the simple comparison of the effectiveness of the various optophonic modifications proposed and applied when it came to the last aspect of DeLIA (namely obstacle avoidance). Tables and chairs were arranged in the form of a basic obstacle course, which the test subjects had to navigate using only the sound output from the optophone system. The subjects were not allowed to reach out and feel for the location of an object. Within the arrangement of furniture there existed a

gap between two tables of approximately one metre, which the subjects were required to locate and manoeuvre through.

This experiment was designed to provide an insight into whether the proposed optophonic modifications could be used as an aid to recognition and avoidance of everyday items that pose a threat to the user's safety. Four different types of optophonic system were tested, namely the standard video optophone with its plain unmodified images, an edge depth map, the Triclops systems innate depth map, and the cartoon depth map. Of these mappings, subjects found that the plain images and the edge depth map provided almost no additional help. The reasons for this apparent lack of assistance can be found in the similarity in colour of the tables, chairs, and the carpet of the room used, all of which were of various tones of brown (represented as a fairly dark grey by the monochrome cameras used). This rendered recognition more difficult when using the plain images, and would presumably be remedied if all objects within the scene were of significantly different colours.

The edge depth map, although providing accurate depth information, supplied insufficient image detail to correctly determine the location of an object. The image edges enabled the user to find object boundaries, but without touching the object it was not possible to decide on which side of the detected edge the object lay. The only solution to this appeared to be that of moving towards and reaching out for the object's edge, or to spend time tracing out the outline of the object with the video optophone.

A noticeable improvement occurred when using the cartoon depth map as input to the optophonic system, with subjects able to navigate more quickly and easily. However, some difficulties were encountered as the user shifted their gaze. Although the adaptive thresholding worked well in most situations, problems arose when the subject transferred their view from a cluttered scene to one consisting of large areas of limited texture, such as the plain brown carpet. The result appeared to be that the adaptive thresholding would occasionally select the critical value such that some comparatively unremarkable areas within the scene would form filled regions. These filled regions, generally correct in depth, would have the appearance to the user of being an object in their immediate path. Hence, the subject could be confident about

their route between two objects, and then suddenly find a nonexistent obstacle appearing directly ahead. Work is continuing in an attempt to find a more suitable alternative for the adaptive thresholding used in the stereo cartooning to remedy this problem.

Using the intensity based depth map subjects appeared to have few difficulties locating and manoeuvring around the arrangement of obstacles. After a short period of practice with this system subjects were able to decide with some confidence when they came into range of an object, and approximately what distance they were from the object.

5.2.2. Further assessment

The following two tests were used to determine if there is a significant improvement when using the cartoon depth map over the conventional optophonic system.

5.2.2.1. Test3 – Preliminary optophonic evaluation

To further assess whether applying stereo image processing to an optophonic mapping would be beneficial for blind mobility, it was necessary to derive some form of test for use on volunteers. The results of which could be used to refine the system ready for extended evaluation.

5.2.2.1.1. Method

The initial tests consisted of a series of stereo images (figures (5.7a)-(5.7p), pages 138-140) of two large painted letters, A and B, which were placed on the floor of a room at two different distances (0.75m and 1.50m to provide contrasting depths). In total eight sets of stereo images were taken corresponding to the different letter positions (i.e., A – left, B – right, or B – left, A – right, along with the different letter distances), allowing a comparison between user performance with the standard image-to-sound video optophone and any one of the stereo modifications previously

described. Initially the depth map that provided the best visual results, the cartoon depth map, was used for testing. Finally, groups of volunteers were asked, after a ten-minute training period, to identify the position of the letters from the corresponding optophonic sounds. This was designed to test the first three aspects of DeLIA – object detection, location, and identification.

The ten minute training period began with a brief explanation of what each person would hear and what they would be expected to do. They were also told the basics of how the system worked. This included:

- Whilst wearing a pair of headphones, the sounds would appear to travel from their left ear to their right. Consequently, sounds heard mainly in their left ear would be on the left of the image, and vice versa.
- Amplitude (loudness) corresponds to pixel intensity. In other words, for a standard image the darker the image object the louder the sound heard, however, for the depth map, the louder the sound the nearer the object.
- The higher the frequency of the sound heard, the higher position of the corresponding object within the image scene.
- Objects that are distant can be distinguished from closer objects since they appear to have a smaller sound due to the shorter span of time for which they are heard.

It was believed appropriate to assess the performance of the methods under trial on sighted volunteers first. The training session proceeded by displaying the 16 test images (8 displayed both unmodified and as cartoon depth maps – figures (5.7a)-(5.7p)) to the subjects one by one. Each time an image was displayed the equivalent optophonic sound was played three times with an individual run time of two seconds. Between each sound presentation there was a small delay to provide the listener with a period of time to gather their thoughts. Similarly, after the completion of each set of sounds (each image) there was another slightly longer pause before continuing with the next image.

This first stage was found to take just over five minutes, providing sufficient time for the subjects to run through the images again in the order of their choosing so that they could compare and listen for subtle differences. Once the ten minute period was complete they were allowed a short period of time (no more than a minute) to prepare themselves prior to the assessment.

For the test itself the sounds were presented (without the images) in a random order. The subjects were allowed to listen to each sound as many times as they wished until they felt confident about their choice. However, they were not permitted to review previous selections, and they were not allowed to repeat a previously heard scene. Each subject was asked to identify which type of image he or she was listening to (plain or cartoon depth map) and the location of the letters (i.e., 'A' near left, 'B' far right).

The results of these tests are included in Appendix 2 – section (A2.1) and summarised in table (5.4). Table (5.4) lists the subject number and the number of sample repetitions each subject required to complete the test (column heading '**Reps**'). Other columns in the table are headed '**Char**' for the total number of characters correctly determined (whether the subject positioned the 'A' and 'B' in the right location) and '**Depth**' for the total number of character distances determined, as well as '**Total**' for the number of images correctly recognised. A number of the volunteers appeared to tire about three-quarters of the way through the test, and three people commented that they found the continuous switching of sounds between the two techniques (plain image and cartoon depth map) rather trying. Consequently, future tests were devised that used ten samples (randomly repeating two of the eight original images) of only one technique (plain image or depth map) per subject. The programme used in these later tests was standalone (not requiring any additional input other than that provided by the subject), randomly choosing images and recording responses to file.

It should be noted that some errors are visible within the cartoon depth maps found in figures (5.7a)-(5.7p). For instance, obvious errors can be seen in figure (5.7n) to the left of the 'B'. Albeit a problem with the cartoon depth map, after conversion to sound both figure (5.7n) and the original image (figure (5.7m)) portray the same dark region as an area of high amplitude sound. Thus, any errors or artefacts that may appear to occur with the cartoon depth map are no different than that perceived via the unmodified image due to excessively dark regions. This type of error would not have occurred if the region to the left of the 'B' in the original image (figure (5.7m)) was lighter, since cartoon filling only takes place in areas with a dark pixel intensity.

Subject	2.10 Sec							
	Plain Image				Cartoon Depth Map			
	Reps	Char	Depth	Total	Reps	Char	Depth	Total
		Out of 16		Out of 8		Out of 16		Out of 8
1	39	8	10	2	36	12	16	6
2	23	12	7	1	21	12	15	5
3	21	8	9	1	17	6	14	3
4	27	16	11	4	23	14	16	7
5	17	6	9	2	15	8	12	2
6	27	10	10	2	24	10	14	4
7	20	10	14	3	16	12	14	5
Sum	174	70	70	15	152	74	101	32
Average	24.9	10.0	10.0	2.1	21.7	10.6	14.4	4.6

Table (5.4) – Results obtained during initial testing on several volunteer subjects using the images shown in figures (5.7a)-(5.7p).

5.2.2.1.2. Results & analysis

From table (5.4) it can be seen that there is a small reduction in required sample repetition when using cartoon depth maps rather than unmodified images. This implies the cartoon technique is easier and quicker to learn since the recognition rate is faster. There is also an increase in accuracy of depth perception. However, the recognition rate of characters appears almost identical between the two methods used. Overall, when using stereo cartooning, there is an average improvement in accuracy of around 31% (= average increase in image recognition / total number of images = 2.5/8).

Figures (5.8a)-(5.8d) demonstrate the performance of the test subjects with both the cartoon depth maps and the plain, unmodified images. From these figures it can be seen that improvements are made when using cartoon depth maps rather than plain images, with slightly fewer repetitions required for each test and an overall increase in recognition rate for the images due to the enhanced depth perception. However, from figure (5.8b) the recognition rate for the characters (determining which way round they are) reveals no noticeable improvement, and in some cases the cartoon depth map demonstrates a reduction in the number of correctly identified characters in relation to plain images.

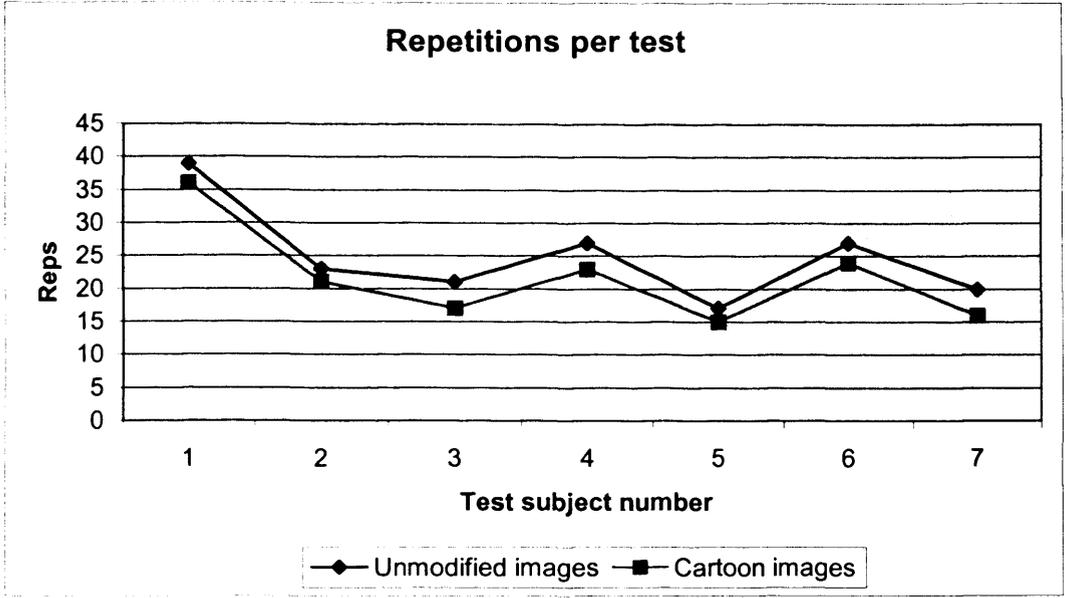


Figure (5.8a) – Graph representing the number of repetitions required by each subject to complete the optophone test with the plain unmodified images and the cartoon depth images.

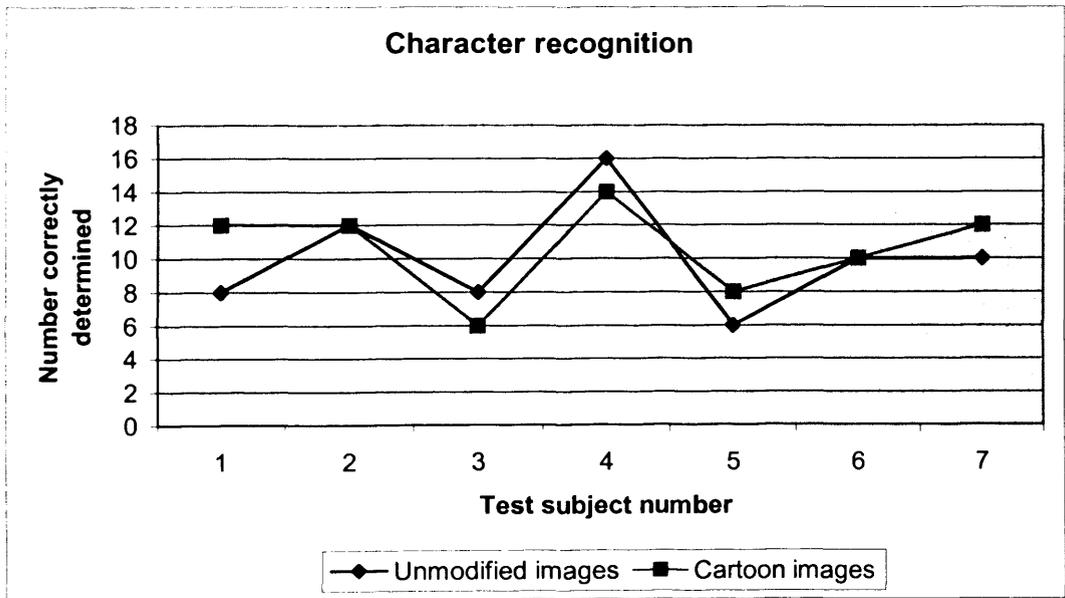


Figure (5.8b) – Graph representing the number of characters correctly identified (out of 16) by each subject during the optophone test with the plain unmodified images and the cartoon depth images.

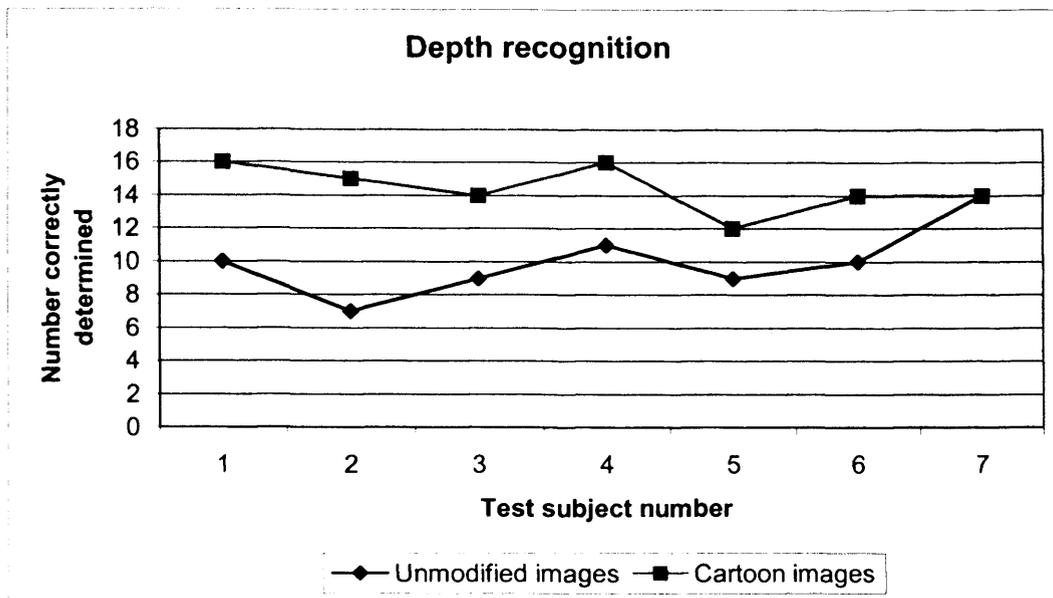


Figure (5.8c) – Graph representing the number of times subjects correctly determined a character’s distance from the observing camera (out of 16) during the optophone test with the plain unmodified images and the cartoon depth images.

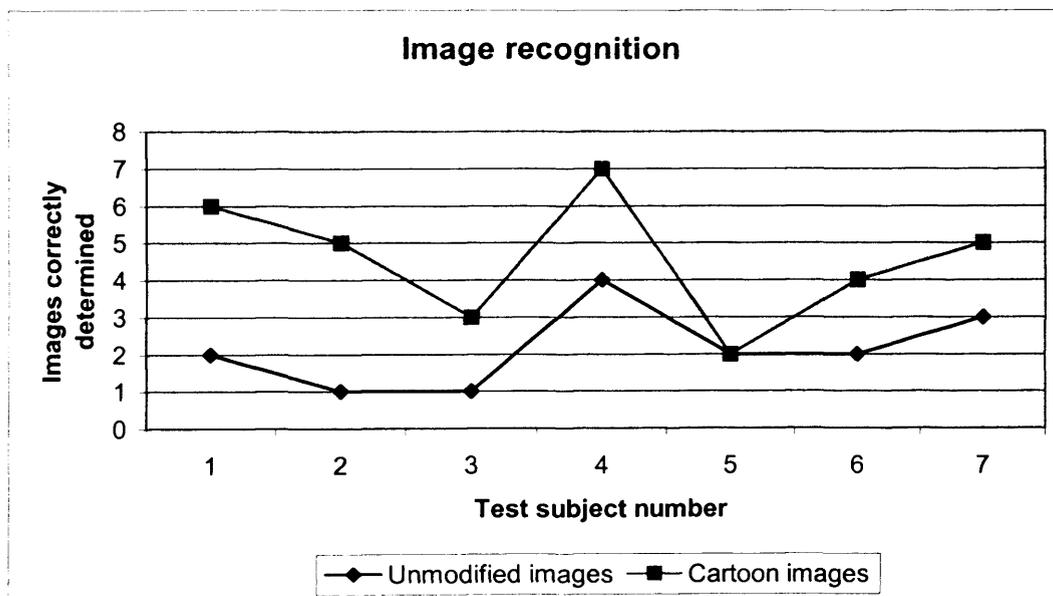


Figure (5.8d) – Graph showing the number of images correctly identified (out of 8) by each subject during the optophone test with the plain unmodified images and the cartoon depth images.

From figures (5.8a)-(5.8d) a general increase can be seen in the recognition rate of test images when using a cartoon depth map rather than a plain image. Whether this difference is statistically significant requires the application of the ‘*paired two sample for means t-test*’ [Cle98, KenCok92]. The *F-test* must be applied to check for homogeneity of the variances of the two samples, which is a prerequisite for the *t-test*. The *F-test* consists of finding the variances of the two sets of data and then dividing the greater by the lesser as shown in equation [5.1].

$$F = \frac{\text{Variance1}}{\text{Variance2}} = \frac{S^2_1}{S^2_2} \quad [5.1]$$

Where $\text{Variance1} > \text{Variance2}$, with the *Variance* being defined as given in equation [5.2].

$$S^2_x = \frac{\sum_{i=1}^N ((X_i - \bar{X})^2)}{N} \quad [5.2]$$

In the case portrayed here, where a sample size of less than 30 values has been used, it is usual to use Bessel’s correction using $N-1$ instead of N (equation [5.3]). However, this alteration changes nothing when calculating the F-value since the $N-1$ cancels during the division.

$$S^2_x = \frac{\sum_{i=1}^N ((X_i - \bar{X})^2)}{N-1} \quad [5.3]$$

If the calculated F value is less than the F-Distribution value (obtained from the F-Distribution table for a predetermined confidence value, i.e., 95%) then the null hypothesis (H_0) that the variances are similar is valid, implying that a parametric test like the *t-test* can be applied to the data sets. If the data is also shown to have a distribution that is approximately normal in nature, as obeyed by the test data (table (5.4)), then another prerequisite of the *t-test* is fulfilled and it can be applied (equation [5.4]).

$$t = \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sum_{i=1}^N (D^2) - \frac{\left(\sum_{i=1}^N D\right)^2}{N}}{N(N-1)}}} \quad [5.4]$$

Where \bar{X} is the mean of the first list of data, and \bar{Y} the second. D is the difference between each pair of X and Y scores, with N representing the number of samples. In equation [5.4] the denominator is sometimes referred to as the ‘Standard Error of the Difference’.

Of the different forms of *t-test* available the one shown in equation [5.4], the ‘*paired two sample for means t-test*’, should be used since each subject provides two items of data for each test performed (indicating that the data is matched/related rather than unrelated). If the absolute (ignoring plus or minus signs) *t-value* calculated from equation [5.4] for the given data is greater than the critical value determined from a lookup table (again at the predetermined level of confidence with $N-1$ degrees of freedom), then the null hypothesis (H_0) is not accepted, which indicates that there is significant difference between the means of the two sets of data. Otherwise the null hypothesis that there is no significant difference between the data sets is accepted.

For instance, consider the number of repetitions required during the tests for both the plain and cartoon depth images given in table (5.4), and shown again in table (5.5). Placing the values obtained from table (5.5) into equations [5.1] & [5.4] reveals the values of F and t as shown in equations [5.5] & [5.6], respectively.

$$F = \frac{312.857/(7-1)}{311.429/(7-1)} = 1.005 \quad [5.5]$$

Subject	Plain Image	Cartoon Depth Map			D	D*D
	Reps	Reps	$(X_i - \bar{X})^2$	$(Y_i - \bar{Y})^2$	(X-Y)	(X-Y)*(X-Y)
	X	Y				
1	39	36	200.020	204.082	3	9
2	23	21	3.449	0.510	2	4
3	21	17	14.878	22.224	4	16
4	27	23	4.592	1.653	4	16
5	17	15	61.735	45.082	2	4
6	27	24	4.592	5.224	3	9
N = 7	20	16	23.592	32.653	4	16
Sum	174	152	312.857	311.429	22	74
Average	$\bar{X} = 24.857$	$\bar{Y} = 21.714$				

Table (5.5) – Data extracted from table (5.4), indicating the number of repetitions required by seven subjects to complete the optophone test, in readiness for performing the *t-test*. The difference ‘D’ and difference squared ‘D*D’ is provided.

In this case the calculated F-value equals 1.005 to three decimal places, which given the critical value of $z=4.28$ obtained from the F-tables with 6df ($N-1$ degrees of freedom) at a 95% confidence level ($F(z)=0.95$), implies that H_0 is valid. Thus, the data allows the use of a parametric test like the *t-test*.

$$t = \frac{|24.857 - 21.714|}{\sqrt{\frac{74 - \frac{(22)^2}{7}}{7(7-1)}}} = \frac{3.143}{0.340} = 9.244 \quad [5.6]$$

With everything taken to three decimal places (3dp), including \bar{X} and \bar{Y} in table (5.5). However, using a software package such as Microsoft Excel to perform the calculations provides an answer of $t=9.242$ (3dp). To determine whether or not this indicates a significant difference between the means of the two values representing the number of repetitions requires a comparison with the relative *t-value*. For 6df ($N-1$) at a significance level of 95% (with the distribution function $F(z)=0.95$) then $z=2.447$ (for a two-tailed test). Therefore, H_0 can be rejected since $|t|>z$, and so there is a significant difference between the mean number of repetitions required for the plain images and cartoon depth maps used during the optophone tests.

Using Microsoft Excel to calculate and display the results for the *t*-test upon the data shown in table (5.4), the values shown in table (5.6) were produced.

t-Test: Paired Two Sample for Means								
	Reps		Char		Depth		Total	
	X	Y	X	Y	X	Y	X	Y
Mean	24.857	21.714	10.000	10.571	10.000	14.429	2.143	4.571
Variance	52.143	51.905	10.667	7.619	4.667	1.952	1.143	2.952
F value	1.005		1.400		2.390		2.583	
F Critical value (95%)	4.284		4.284		4.284		4.284	
Observations (N)	7		7		7		7	
Hypothesised Mean Difference (H_0 – No significant difference)	0		0		0		0	
Df	6		6		6		6	
t value	9.242		-0.679		-4.673		-4.599	
P(T<=t) two-tail (%)	99.991		47.776		99.658		99.630	
t Critical value two-tail (95%)	2.447		2.447		2.447		2.447	
F value H_0 accepted (can use parametric test, $F < z$)	TRUE		TRUE		TRUE		TRUE	
t value H_0 rejected (significant difference, $ t > z$)	TRUE		FALSE		TRUE		TRUE	

Table (5.6) – Results obtained after applying both the *F*-test & *t*-test to the data shown in table (5.4), which was gathered during preliminary tests with the optophone programme.

From table (5.6), all aspects, apart from character recognition, show a significant difference at a 95% confidence level between the mean values acquired for the optophone test when using cartoon depth maps and plain images. In fact the percentage values given in the table corresponding to ‘P(T<=t) two-tail’ represent the maximum confidence level at which the results reveal a significant difference.

Now that a statistical difference has been shown it is important to determine with which technique an improvement is demonstrated (i.e., whether subjects perform better with plain images or cartoon depth maps). This is achieved by looking at the mean values for pairs of data from table (5.4), revealing that the cartoon depth maps

demonstrate an improvement over the plain images in all three of the cases in which the *t-test* indicates a significant difference.

Although this experiment cannot adequately supply information as to the subjectively perceived effects on the user, such as how tiresome the particular techniques were, all subjects commented that the sounds resulting from the cartoon images were easier to listen to since they appeared quieter and less penetrating. This is due to the depth maps having large blank areas caused by the adaptive cartoon thresholding that has been designed to fill no more than a specific quantity of the image. For the test images, the adaptive threshold was set so that the algorithm chose a pixel intensity limit corresponding to 30% or less of the image being filled (>70% blank – white pixels). When generating the optophonic sound these blank areas are noiseless, whereas with the unmodified images all frequencies are usually present to varying degrees. Further improvements to the often-overwhelming sound are also brought about by the cartoon depth map through the varying amplitude caused by the differing distance of objects. The more distant the detected object, the quieter the corresponding optophonic sound.

5.2.2.2. Test4 – Further optophonic testing

The results obtained for the preliminary testing, as shown in the previous section, were very encouraging. Before proceeding to a larger test sample it was necessary to calculate the sample size that would detect a specified difference (if it existed) with an appropriate level of confidence. The comparison of matched pairs for means tests the null hypothesis (H_0) that the mean difference between matched pairs is equal to zero versus the alternative two-tailed hypothesis (H_A) that the mean difference does not equal zero. Thus, to estimate the sample size required to conclude that the mean difference is significantly different from zero with a probability of type I error, the formula given by equation [5.7] can be used, which is similar to that for the confidence interval for the mean. [DesRag90]

$$np = \frac{(SD)^2 \times z_{\alpha}^2}{mdiff^2} \quad [5.7]$$

The total sample size for a two-tailed test is then defined in equation [5.8].

$$N = 2 \times np \quad [5.8]$$

Where np is the number of matched pairs; N is the total sample size; SD is the standard deviation of the differences; and $mdiff$ the mean difference between matched pairs. With $alpha$ being the probability of a type I error, which in this case corresponds to 0.050 (for a 95% confidence level or 5% error), and $zalpha$ being the corresponding z-score for alpha of 1.960.

Setting $mdiff$ equal to 1.000, which implies an expected mean difference of at least one between the performance of subjects with the plain images and cartoon depth images with all aspects tested. With a standard deviation of the differences of 0.900 for the number of repetitions (estimated from the preliminary test data), then np equals 3.112 and the total number of samples required to determine whether the mean difference is significant is $N = 6.223$ 3dp. Thus, only 7 samples are required (N rounded up to the nearest integer).

Performing the same tests on the data for the characters recognised ($SD = 2.225$), depth correctly determined ($SD = 2.507$), and the total images recognised ($SD = 1.397$) gives N equal to 38.037, 48.289, and 14.995, respectively. However, for the characters recognised and the depth correctly determined (which have a possible score out of 16) it would be expected that the calculations be performed using a mean difference twice as high as that for the number of images correctly determined, which only has a score out of 8. Taking $mdiff$ as equal to 2 provides a sample size N of 9.509 and 12.072 for the characters and depth recognition, respectively.

From the reasoning described above, the test was modified and performed on another larger group of 20 subjects (since this quantity of test subjects exceeded the required sample sizes derived from equations [5.7] & [5.8]), the results of which can be found in Appendix 2 – section (2.2) and summarised in table (5.7).

The programme used for testing (screenshots shown in Appendix 3 – section (A3.2.2.2)) was redesigned to provide the subject with a two-minute practice period

followed by the test. The test consisted of two parts: first the subject's performance was evaluated with one type of image, then after a suitable break the subject was tested with the second class of image. Each new subject performed the test in reverse order to the previous in an attempt to be non-biased to any particular type of image.

In each session the subject heard a random set of ten different sounds, the eight test images of the particular type being tested with two additional repeats. The results obtained from these tests are shown in table (5.7), and represented graphically in figures (5.9a)-(5.9d).

Subject	Test order	2.10 Sec							
		Plain images (P)				Cartoon depth map (C)			
		Reps	Char	Depth	Total	Reps	Char	Depth	Total
			Out of 20		Out of 10		Out of 20		Out of 10
1	CP	25	14	9	1	24	12	19	6
2	PC	24	12	13	2	21	10	20	5
3	CP	49	16	16	7	42	14	17	5
4	PC	16	20	17	7	13	18	20	8
5	CP	28	10	14	2	23	12	18	6
6	PC	27	14	11	3	24	16	14	4
7	CP	31	10	10	1	31	8	16	3
8	PC	14	8	12	1	11	16	15	5
9	CP	13	10	12	2	12	8	14	3
10	PC	24	12	8	1	18	10	17	4
11	CP	18	10	9	0	17	12	19	5
12	PC	31	14	10	2	25	6	15	1
13	CP	17	12	10	2	18	6	14	2
14	PC	25	10	10	1	25	8	16	2
15	CP	33	18	12	3	43	10	15	4
16	PC	50	12	8	1	31	10	16	4
17	CP	53	6	10	0	36	12	15	3
18	PC	48	10	10	2	26	6	16	1
19	CP	29	10	10	1	32	10	14	1
20	PC	32	14	13	2	25	14	17	4
Sum		587	242	224	41	497	218	327	76
Average		29.35	12.10	11.20	2.05	24.85	10.90	16.35	3.80

Table (5.7) – Results obtained during testing on twenty volunteer subjects using the images shown in figures (5.7a)-(5.7p).

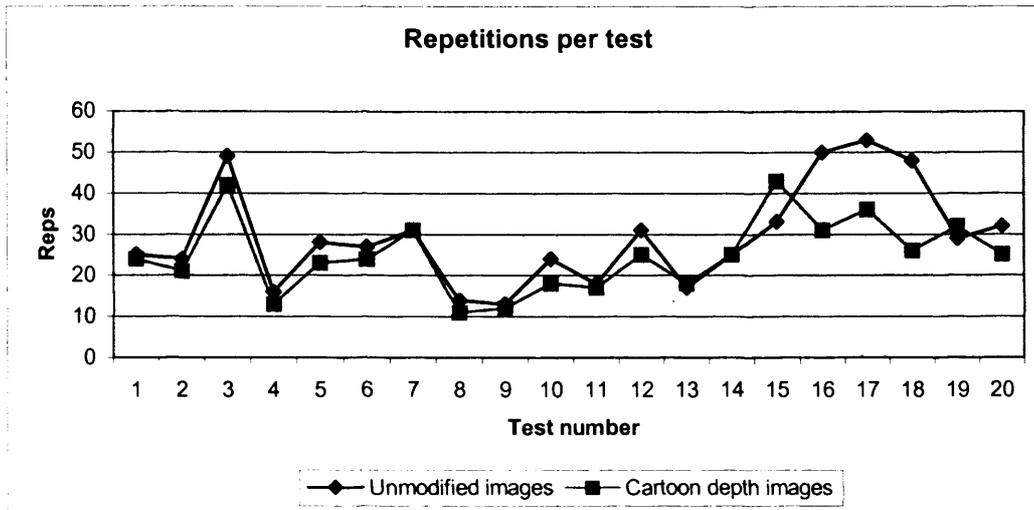


Figure (5.9a) – Graph representing the number of repetitions required by each of the twenty subjects to complete the optophone test with the plain unmodified images and the cartoon depth images.

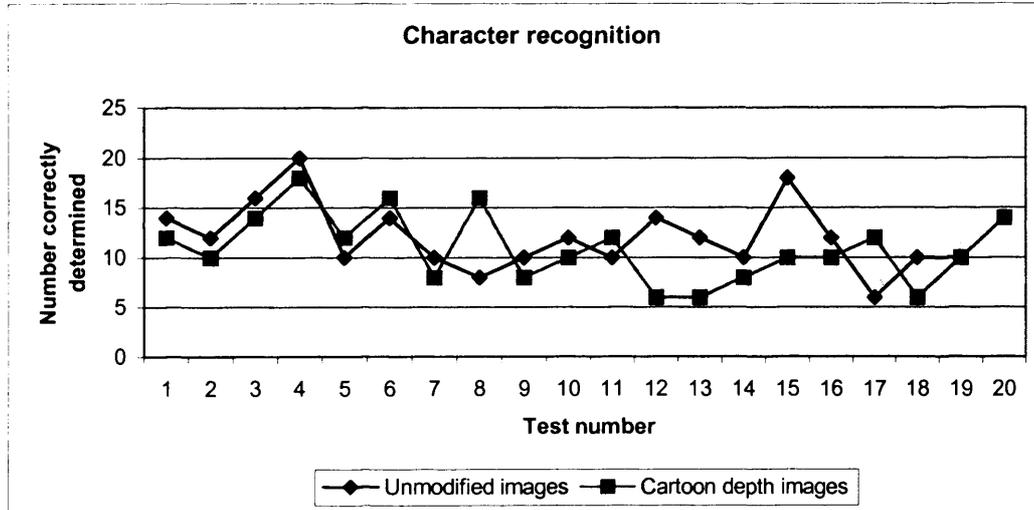


Figure (5.9b) – Graph representing the number of characters correctly identified (out of 20) by each subject during the optophone test with the plain unmodified images and the cartoon depth images.

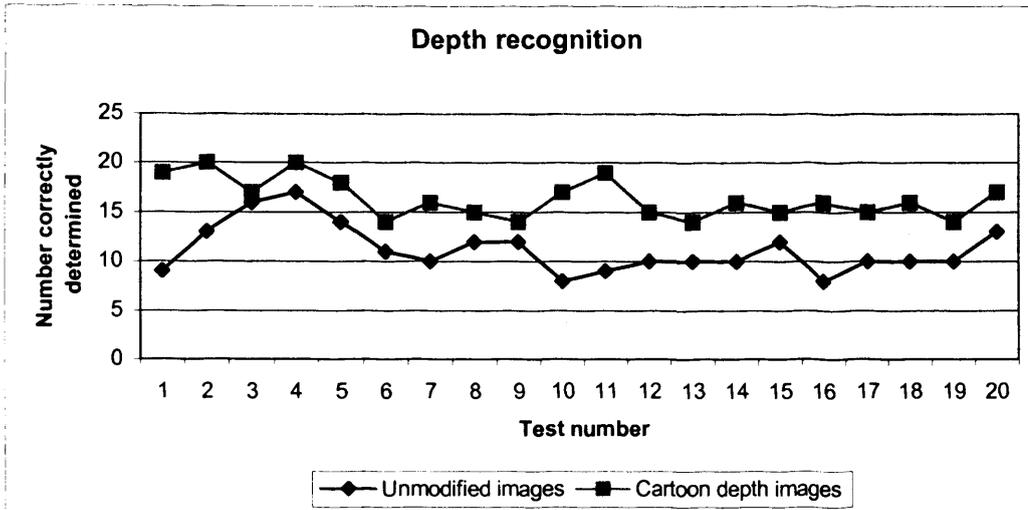


Figure (5.9c) – Graph representing the number of times subjects correctly determined a character’s distance from the observing camera (out of 20) during the optophone test with the plain unmodified images and the cartoon depth images.

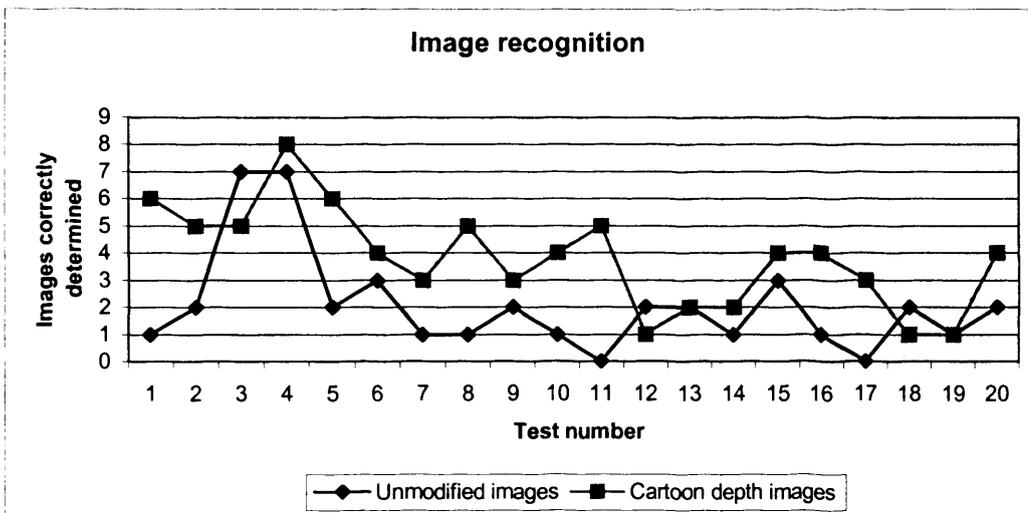


Figure (5.9d) – Graph showing the number of images correctly identified (out of 10) by each subject during the optophone test with the plain unmodified images and the cartoon depth images.

As with the first test of this nature (section 5.2.2.1), from figures (5.9a)-(5.9d) it is possible to see a general improvement in recognition of test images when using a cartoon depth map rather than a plain image. The only feature that the cartoon depth

map demonstrates no improvement with is character identification. In this case the difference, albeit very small, appears to favour the plain images on more occasions than the depth maps. However, whether there is actually a significant difference in obtainable performance between the two types of image can only be shown by the application of the 'paired two sample for means t-test', described in section (5.2.2.1.2). Calculating the *F-value* and *t-value* for the data shown in table (5.7) provides the results given in table (5.8).

t-Test: Paired Two Sample for Means								
	Reps		Char		Depth		Total	
	X	Y	X	Y	X	Y	X	Y
Mean	29.350	24.850	12.100	10.900	11.200	16.350	2.050	3.800
Variance	147.924	81.503	11.147	11.989	5.958	3.924	3.524	3.432
F value	1.815		1.076		1.518		1.027	
F Critical value (95%)	2.168		2.168		2.168		2.168	
Observations (N)	20		20		20		20	
Hypothesised Mean Difference (H0 - No significant difference)	0		0		0		0	
Df	19		19		19		19	
t value	2.678		1.352		-8.924		-3.920	
P(T<=t) two-tail (%)	98.512		80.786		100.000		99.908	
t Critical value two-tail (95%)	2.093		2.093		2.093		2.093	
F value H ₀ accepted (can use parametric test, F<z)	TRUE		TRUE		TRUE		TRUE	
t value H ₀ rejected (significant difference, t >z)	TRUE		FALSE		TRUE		TRUE	

Table (5.8) – Results obtained after applying both the *F-test* & *t-test* to the data shown in table (5.7), which was gathered during the second set of tests with the optophone programme. Calculated values are shown to three decimal places.

As before with the preliminary results (table (5.6)) it can be seen from table (5.8) that all aspects, apart from the character recognition, show a significant difference at a 95% confidence level (and higher, as indicated by the percentage value 'P(T<=t) two-tail' in the table) between the mean values acquired for the optophone test when using cartoon depth maps and plain images.

From the results in table (5.8) the direction of significant difference can be ascertained by comparing the mean values for each of the four aspects tested and found to have a *true* difference by the *t-test*. This reveals that the cartoon depth maps demonstrate an improvement over the plain images with respect to the quantity of repetitions per test, and the number of correctly determined depth estimates as well as the image recognition rate.

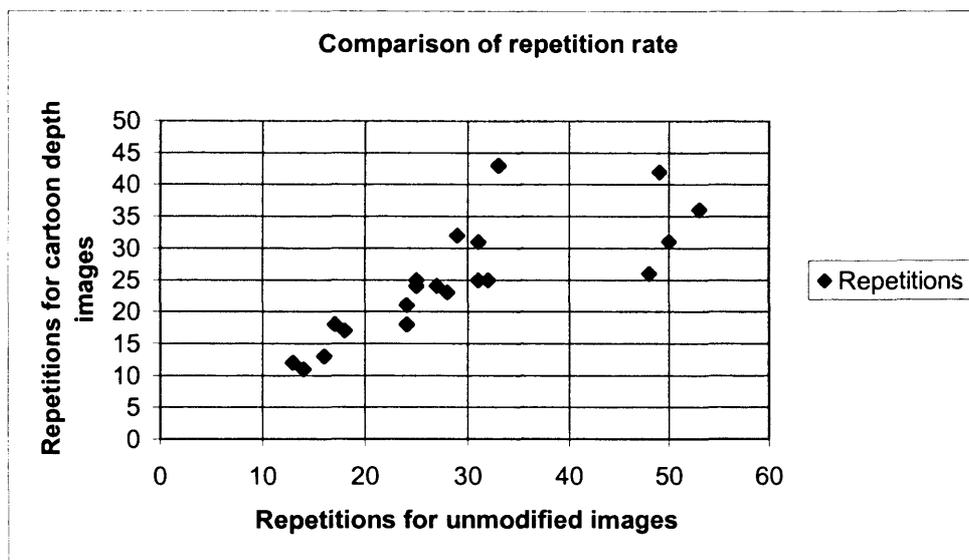


Figure (5.10) – A scatter graph comparing the number of required repetitions to complete the optophone test with the unmodified images along the x-axis with those for the cartoon depth images along the y-axis.

It may be noticed that the maximum confidence level for the *t-test* at which the quantity of repetitions shows a significant difference between the two types of image has deteriorated slightly in comparison to the results obtained for the initial seven test subjects (table (5.6)). The original calculation gave a maximum limit of 99.991% (3dp), whereas the later test provided a value of 98.512% (3dp). By looking at a scatter graph of the results (figure (5.10)) for the twenty test subjects the reason for this drop in value can be seen. Several outliers exist (situated about the line $x=50$, where x corresponds to the *unmodified images*), which although in favour of the cartoon depth map, reduce what would otherwise be a highly correlated set of results. Consequently a reduction is observed in the maximum confidence level of the *t-test*.

5.3. Summary

During the research into various methods of stereo processing it appeared that no suitable automated system existed for the evaluation of depth maps, other than that provided by visual inspection. The techniques that do exist for these purposes generally involve comparing the results of a computer generated depth map with one created by hand. A job that is both laborious and time consuming, often requiring a great deal of patience. Hence, a system was derived that, although not perfect, could provide a reliable measure as to the efficiency of a stereo algorithm. The method sets a disparity range greater than required for the chosen evaluation images, then compares known false matches with assumed correct matches. Although an assumption is made, the test is relatively reliable since the same assumption is made for each algorithm under test. Bearing in mind that the results obtained are not meant as an accurate ratio of correct to incorrect matches, but rather as a gauge of the likelihood of the stereo algorithm to make incorrect matches. In other words, it is a gauge of the number of points known to have been incorrectly matched within a certain predefined disparity range, or the FPF (False Positive Fraction).

The most suitable stereo edge algorithm, which was combined with the stereo cartooning technique, was determined both via visual means and with the FPF method for evaluating the efficiency of stereo processes. Tests were then derived to assess whether stereo cartooning would be beneficial to an optophonic mapping.

The proposed method of evaluation was labelled DeLIA for the Detection, Location, Identification and Avoidance (or Action) of obstacles, representative of the process that a blind user might use to navigate through an unfamiliar environment using a mobility aid. Tests were formed using volunteers in an effort to assess each of the stages involved in DeLIA.

An experiment was performed, which although subjective in nature, demonstrated that subjects could perceive basic images at extreme frame rates (32 frames per second),

implying that with training the brain might adapt to process greater quantities of information via sound.

Using the optophone with various stereo depth maps it was shown that subjects could locate and manoeuvre around potential obstacles. Although the cartoon depth map showed promise with respect to this task, minor problems with the adaptive thresholding caused some spurious results. This occurred mainly whilst moving the view field around areas of limited texture, such as a plain carpet.

The intensity depth map provided by the Triclops system, albeit unable to display text, was found to provide suitable image detail to allow subjects to perceive objects placed within a basic obstacle course designed to assess the final stage of DeLIA (obstacle Avoidance).

Further, more rigorous tests consisted of a series of images of two large painted letters, A and B, which were placed on the floor of a room at two different distances. In total eight images were taken corresponding to the different combinations of letter position. The equivalent cartoon depth maps were generated from these images, allowing a comparison between the two techniques. Finally, groups of volunteers were asked, after a short training period, to identify the position of the letters from the corresponding optophonic sounds. The experiments revealed that pre-processing the optophonic input with the method of stereo cartooning improved the perception of depth, and lessened both the number of repetitions required to recognise the samples and the stress with which the user perceived the sounds.

The experiments also indicated that the ideal input to the optophonic mapping is one that is capable of providing the user with similar scene information as to that perceived by a sighted person (through the visual system), but with a considerable reduction in emphasis of unimportant features (in this case, provided by objects that are beyond the range of the user).

Figure (5.1a)-(5.1z) - Lamp-Office-ABC Image Set - Stereo Algorithm Test

A set of depth maps generated from the various stereo algorithms that were tested whilst searching for the technique that provides the best performance in terms of speed, accuracy, and visual appearance. The values provided in the figure captions correspond to results obtained using the FPF evaluation technique described in section (5.1.1) and listed in table (5.2), page 104.

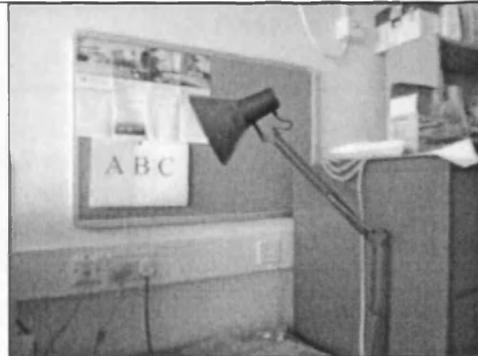


Figure (5.1a) – Left stereo image. One of the stereo samples used to evaluate the performance of various stereo algorithms via the quantity of False Positives.

Figure (5.1b) – Right stereo image of an office lamp. Figures (5.1a) & (5.1b) have dimensions 320x240 pixels.

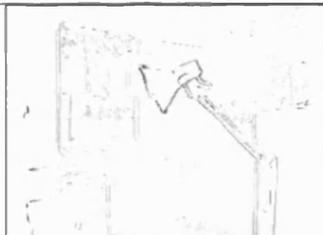


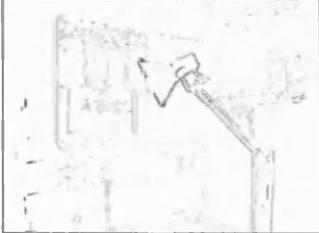
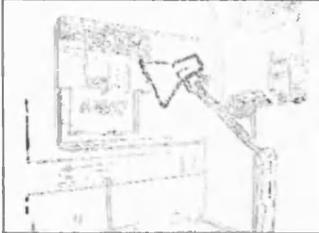
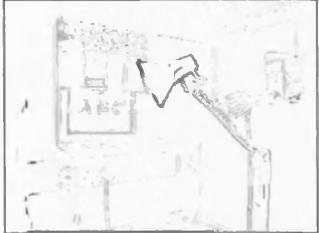
Figure (5.1c) – Combination 1 – A0R0C0S0T0 – 52% of edges assumed to be correctly matched at 9.1 frames/second.

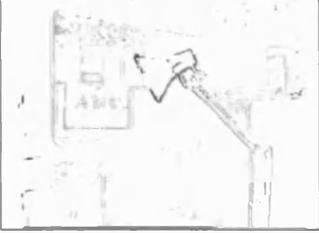
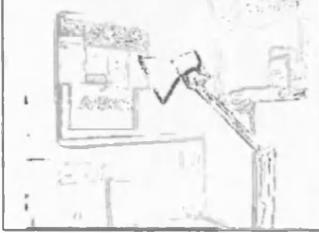
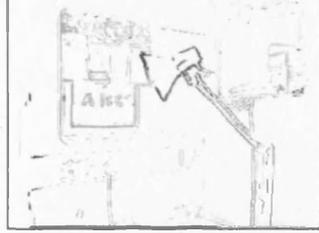


Figure (5.1d) – Combination 2 – A0R0C0S0T1 – Matched 17% of edges at 3.7 frames/second.



Figure (5.1e) – Combination 3 – A0R0C0S1T0 – Matched 36% of edges 3.6 frames/second.

		
<p>Figure (5.1f) – Combination 4 – A0R0C0S1T1 – 46% at 1.7 frames/sec.</p>	<p>Figure (5.1g) – Combination 5 – A0R0C1S0T0 – 52% at 9.5 frames/sec.</p>	<p>Figure (5.1h) – Combination 6 – A0R0C1S0T1 – 19% at 4.1 frames/sec.</p>
		
<p>Figure (5.1i) – Combination 7 – A0R0C1S1T0 – 36% at 5.0 frames/sec.</p>	<p>Figure (5.1j) – Combination 8 – A0R0C1S1T1 – 46% at 2.0 frames/sec.</p>	<p>Figure (5.1k) – Combination 9 – A0R1C0S0T0 – 57% at 9.1 frames/sec.</p>
		
<p>Figure (5.1l) – Combination 10 – A0R1C0S0T1 – 49% at 3.2 frames/sec.</p>	<p>Figure (5.1m) – Combination 11 – A0R1C0S1T0 – 33% at 5.0 frames/sec.</p>	<p>Figure (5.1n) – Combination 12 – A0R1C0S1T1 – 41% at 1.9 frames/sec.</p>

		
<p>Figure (5.1o) – Combination 13 – A0R1C1S0T0 – 57% at 9.5 frames/sec.</p>	<p>Figure (5.1p) – Combination 14 – A0R1C1S0T1 – 50% at 3.6 frames/sec.</p>	<p>Figure (5.1q) – Combination 15 – A0R1C1S1T0 – 33% at 5.4 frames/sec.</p>
		
<p>Figure (5.1r) – Combination 16 – A0R1C1S1T1 – 40% at 2.2 frames/sec.</p>	<p>Figure (5.1s) – Combination 17 – A1R0C0S0T0 – 63% at 8.7 frames/sec.</p>	<p>Figure (5.1t) – Combination 18 – A1R0C0S0T1 – 20% at 3.5 frames/sec.</p>
		
<p>Figure (5.1u) – Combination 19 – A1R0C1S0T0 – 63% at 9.1 frames/sec.</p>	<p>Figure (5.1v) – Combination 20 – A1R0C1S0T1 – 21% at 4.0 frames/sec.</p>	<p>Figure (5.1w) – Combination 21 – A1R1C0S0T0 – 65% at 8.7 frames/sec.</p>

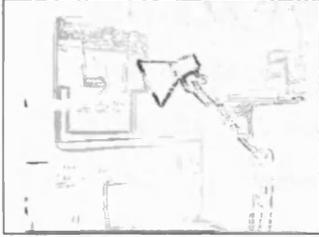


Figure (5.1x) – Combination 22 – A1R1C0S0T1 – 48% at 3.2 frames/sec.

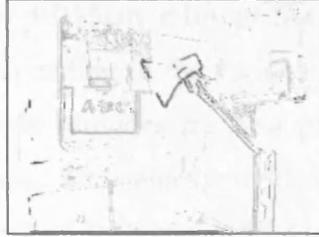


Figure (5.1y) – Combination 23 – A1R1C1S0T0 – 64% at 9.1 frames/sec.

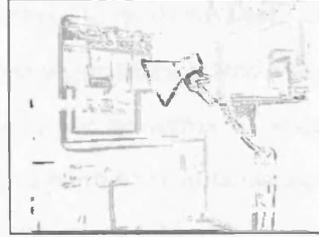


Figure (5.1z) – Combination 24 – A1R1C1S0T1 – 48% at 3.6 frames/sec.

Figure (5.2a)-(5.2z) - A150cm_l-B75cm_r Image Set - Stereo Algorithm Test

A second set of figures corresponding to the various stereo algorithms tested whilst searching for the technique that provides the best performance in terms of speed, accuracy, and visual appearance. The values provided correspond to results obtained using the FPF technique described in section (5.1.1) and listed in table (5.3), page 105.



Figure (5.2a) – Left stereo image. Another one of the stereo samples used in evaluating the performance of various stereo algorithms via the quantity of False Positives. Both this and figure (5.2b) have dimensions of 320x240 pixels.

Figure (5.2b) – Right stereo image. In the above scene the ‘A’ is 1.50 metres from the camera, and the ‘B’ is 0.75 metres. The greatest disparity (from the ‘B’) was found to be 36 pixels, and the smallest was 11 pixels.

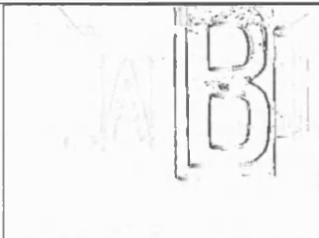


Figure (5.2c) – Combination 1 – A0R0C0S0T0 – 44% of edges assumed to be correctly matched at 8.3 frames/second.

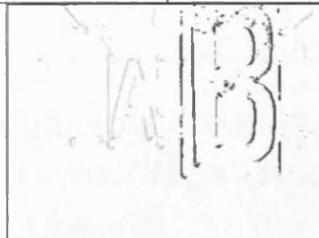


Figure (5.2d) – Combination 2 – A0R0C0S0T1 – Matched 16% of edges at 3.7 frames/second.

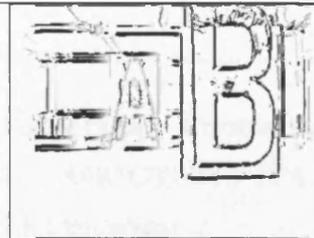


Figure (5.2e) – Combination 3 – A0R0C0S1T0 – Matched 33% of edges 3.9 frames/second.

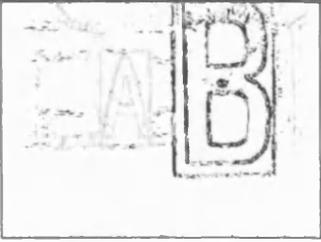


Figure (5.2f) – Combination 4 – A0R0C0S1T1 – 40% at 1.6 frames/sec.

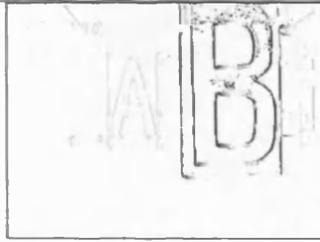


Figure (5.2g) – Combination 5 – A0R0C1S0T0 – 44% at 9.1 frames/sec.



Figure (5.2h) – Combination 6 – A0R0C1S0T1 – 15% at 4.3 frames/sec.

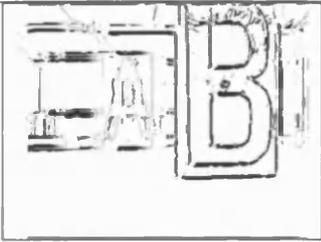


Figure (5.2i) – Combination 7 – A0R0C1S1T0 – 34% at 4.2 frames/sec.

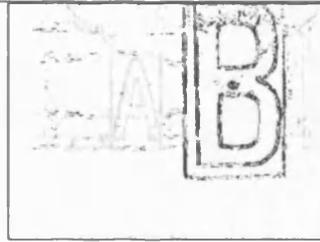


Figure (5.2j) – Combination 8 – A0R0C1S1T1 – 40% at 1.9 frames/sec.

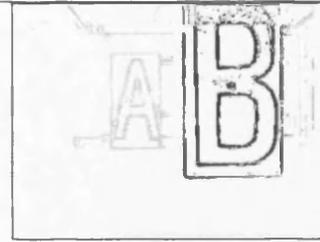


Figure (5.2k) – Combination 9 – A0R1C0S0T0 – 58% at 8.3 frames/sec.

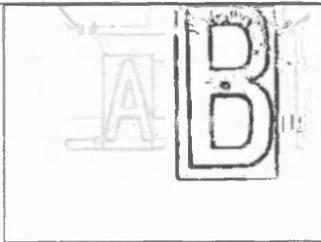


Figure (5.2l) – Combination 10 – A0R1C0S0T1 – 51% at 3.1 frames/sec.

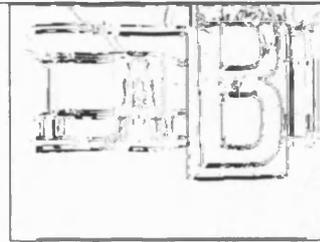


Figure (5.2m) – Combination 11 – A0R1C0S1T0 – 28% at 4.4 frames/sec.

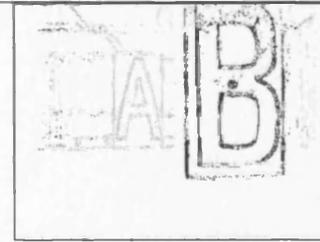


Figure (5.2n) – Combination 12 – A0R1C0S1T1 – 35% at 1.8 frames/sec.

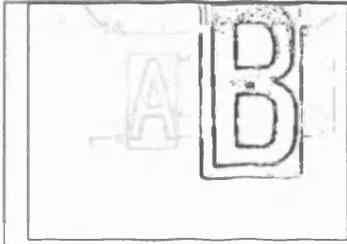


Figure (5.2o) – Combination 13 – A0R1C1S0T0 – 57% at 8.7 frames/sec.

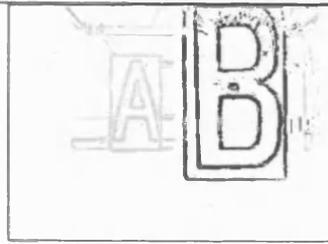


Figure (5.2p) – Combination 14 – A0R1C1S0T1 – 51% at 3.5 frames/sec.

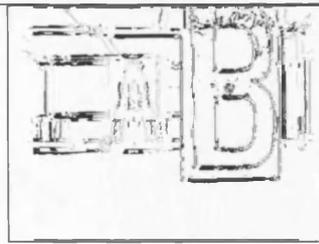


Figure (5.2q) – Combination 15 – A0R1C1S1T0 – 28% at 4.9 frames/sec.

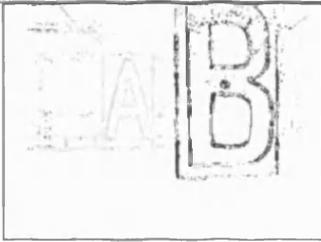


Figure (5.2r) – Combination 16 – A0R1C1S1T1 – 35% at 2.2 frames/sec.

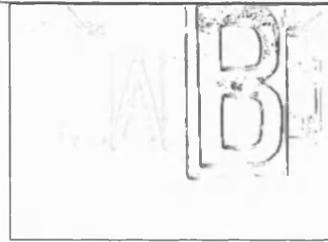


Figure (5.2s) – Combination 17 – A1R0C0S0T0 – 60% at 8.0 frames/sec.

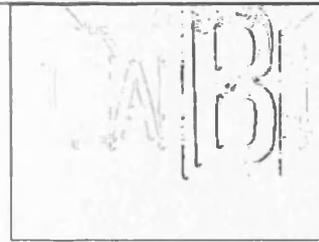


Figure (5.2t) – Combination 18 – A1R0C0S0T1 – 17% at 3.6 frames/sec.

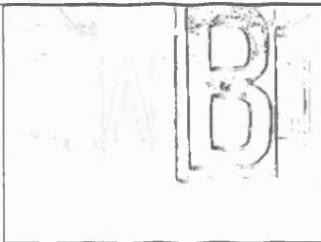


Figure (5.2u) – Combination 19 – A1R0C1S0T0 – 59% at 8.3 frames/sec.

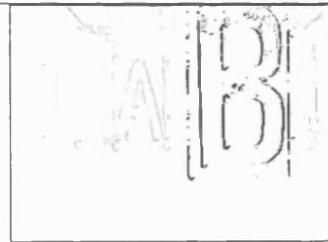


Figure (5.2v) – Combination 20 – A1R0C1S0T1 – 17% at 4.0 frames/sec.

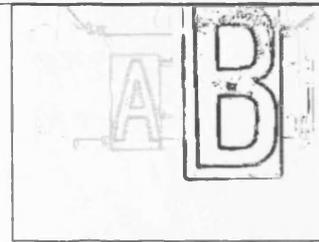


Figure (5.2w) – Combination 21 – A1R1C0S0T0 – 68% at 8.0 frames/sec.

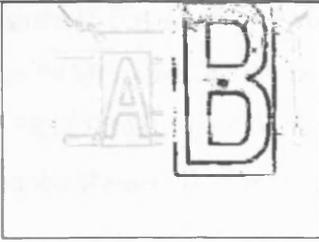


Figure (5.2x) – Combination
22 – A1R1C0S0T1 – 51% at
3.1 frames/sec.

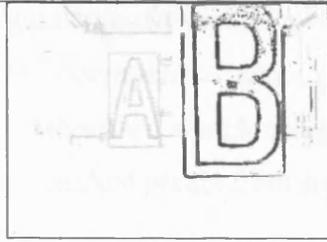


Figure (5.2y) – Combination
23 – A1R1C1S0T0 – 68% at
8.3 frames/sec.

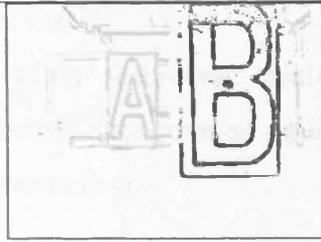


Figure (5.2z) – Combination
24 – A1R1C1S0T1 – 51% at
3.5 frames/sec.

Figure (5.5a)-(5.5b) - PMF algorithm - Stereo Algorithm Test

The PMF constraint used in conjunction with the stereo technique labelled 'A1R1C0S0T0' (indicating the **Advanced** search using previous disparities as a guide and the **Removal** of previously matched pixels from future searches).

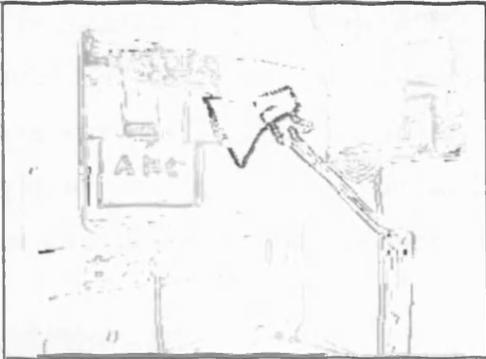


Figure (5.5a) – The PMF algorithm applied to figures (5.1a) & (5.1b), page 129.

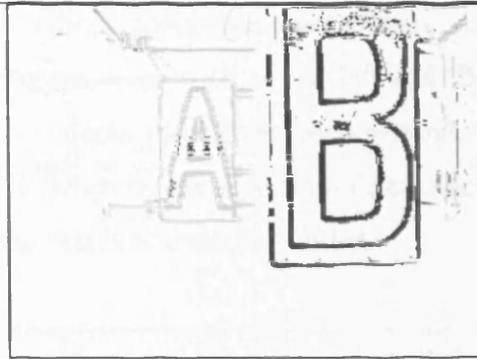


Figure (5.5b) – The PMF algorithm applied to figures (5.2a) & (5.2b), page 133.

Figure (5.7a)-(5.7p) – Sample images used for evaluating optophonic techniques with DeLIA.

Set of figures consisting of eight images showing two letters ('A' & 'B') painted onto two large upright boards. Each image represents one photo taken from a stereo set of three. The other eight images shown correspond to cartoon depth maps generated from the stereo sets. The letter boards shown have been positioned corresponding to distances of 0.75m and 1.50m, as well as differing positions such as 'A' left and 'B' right, or vice versa. All figures portraying cartoon depth maps have been generated using an adaptive cartoon-filling threshold of 30%, whereby the algorithm determines and uses a critical value that generates a depth map that is at most 30% filled.



Figure (5.7a) – 'A' 1.50m & 'B' 1.50m.

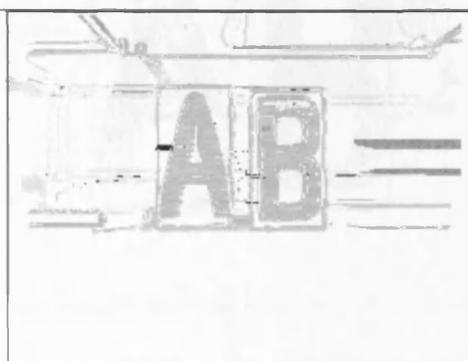


Figure (5.7b) – Cartoon depth map generated from image set figure (5.7a).



Figure (5.7c) – 'A' 1.50m & 'B' 0.75m.

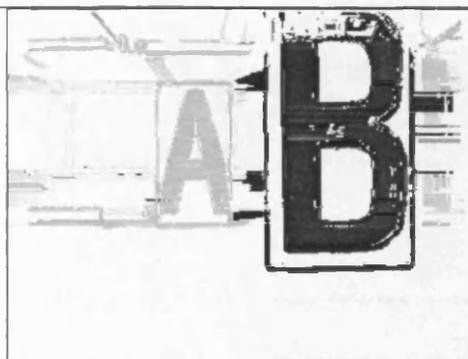


Figure (5.7d) – Cartoon depth map generated from image set figure (5.7c).



Figure (5.7e) – ‘A’ 0.75m & ‘B’ 1.50m.

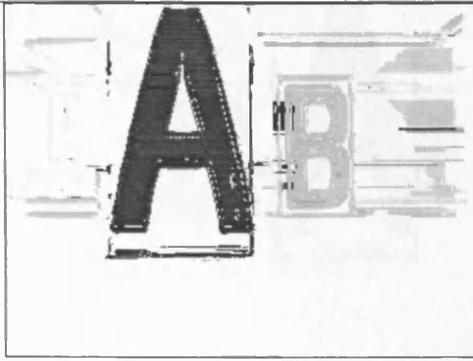


Figure (5.7f) – Cartoon depth map generated from image set figure (5.7e).



Figure (5.7g) – ‘A’ 0.75m & ‘B’ 0.75m.



Figure (5.7h) – Cartoon depth map generated from image set figure (5.7g).



Figure (5.7i) – ‘B’ 1.50m & ‘A’ 1.50m.

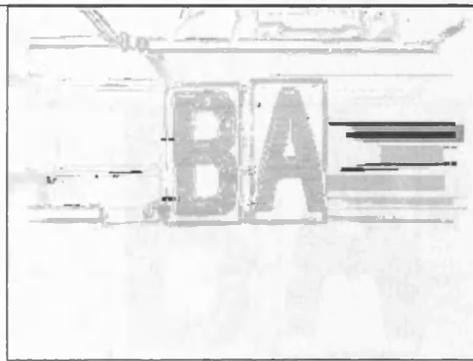


Figure (5.7j) – Cartoon depth map generated from image set figure (5.7i).



Figure (5.7k) – ‘B’ 1.50m & ‘A’ 0.75m.

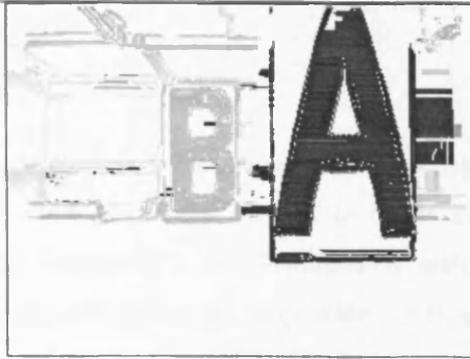


Figure (5.7l) – Cartoon depth map generated from image set figure (5.7k).



Figure (5.7m) – ‘B’ 0.75m & ‘A’ 1.50m.

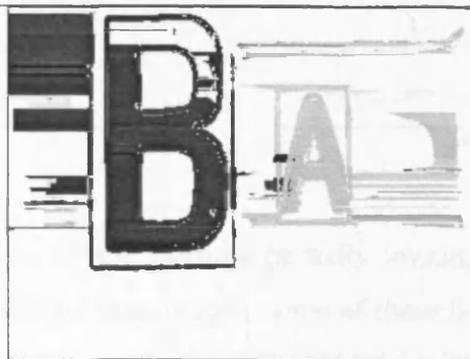


Figure (5.7n) – Cartoon depth map generated from image set figure (5.7m). Some obvious errors have occurred in this image to the left of the ‘B’.



Figure (5.7o) – ‘B’ 0.75m & ‘A’ 0.75m.



Figure (5.7p) – Cartoon depth map generated from image set figure (5.7o).

6. Other approaches

This section describes numerous areas that were provisionally investigated, and that hopefully will be scrutinised more thoroughly in future work, but were deemed to be of a lesser importance at the time. Amongst the topics discussed is the fundamental matrix for calculating the stereo epipolar line, and the use of 3-D soundscapes to provide a listener with a better representation of an object's location. Other ideas considered include the use of a single fixed focused camera to provide images of a scene that are subsequently filtered with a number of different sized edge operators. Object distances being determined by the quantity of edge blurring.

During the research there were many topics, which although partially investigated, were left for future development due to inevitable constraints. Some of these lines of investigation may, if pursued, produce promising results, as indicated by the findings of the preliminary research. In following sections the more interesting and promising topics are discussed.

6.1. 3-D soundscapes

Although relatively untouched during the research in favour of methods of image processing, 3-D soundscapes are one such area that could significantly aid in the representation of a depth map. In a three-dimensional soundscape each object in the real world scene generates a constant individual sound. Each encoded so the listener perceives the sound as originating from the object itself. This area of research, used in conjunction with stereo depth maps, could constitute a major advancement in mobility aids requiring a separate research project.

6.2. The Neurophone

Whilst considering possible difficulties encountered by an optophone user it was realised that although partial success may be made in compensating for loss of sight, it would be made at a cost. The optophone's output considered during the research was stereo sound, presented to the user via headphones. Unfortunately, for the blind, auditory information is one of the most important factors for safety during mobility, since not only can it be used to provide echolocation clues (footfalls echoing of nearby walls), it also indicates the presence of nearby hazards (noise from a busy road). When using the optophone many of these important details gained through the auditory channel are lost.

During the initial stages of research a possible solution was found that provided an alternative and novel method for presenting sound to the user. Preliminary tests were carried out on a device, called the Neurophone, which it was claimed presented sound to a listener by channels other than the auditory system.

It must be noted that no reference has been provided for the discussion on the Neurophone because various 'changing' sources from the internet were originally used prior to purchase. However, a search performed on the internet for the keyword 'Neurophone' produces many relevant links. It has also been reported that Harry Stine, scientist and author, wrote a book called **The Silicon Gods** (Bantam Books) about the potential of the Neurophone as a brain-to-computer connecting interface device.

In 1958, while only 14 years old, Patrick Flanagan invented the device known as the Neurophone. The device has met a great deal of scepticism and until fairly recently was relatively unheard of. It is even reported (from sources on the internet) that the device has worked on some deaf subjects who volunteered to test the device.

The Neurophone has an audio input that can come from a microphone, radio, etc., and an output that leads to two transducers worn as a headband. The transducer disks are made from Zirconium Titanate embedded in acrylic plastic tiles possessing the same

dielectric constant as human skin. It is believed that due to the skin's piezoelectric and optoelectric nature, the Neurophone signals enter the body through certain nerve endings, travelling up these nerves into the brain.

After only a few minutes of use with the Neurophone the brain appears to tune into the previously undetectable sounds emanating from the device, and consequently, from that point on it can be used with no further difficulties. Although the description seems rather bizarre it appears that only the person holding the electrodes can perceive the sound, and whilst doing so it is possible to readily perceive sounds via normal means. The sound perceived from the Neurophone is best described as a background noise loud enough to be interpreted whilst not interfering with external sounds, analogous to playing background music whilst having a conversation with a group of people. The sound appears to originate from within the centre of the head.

Using a device such as the Neurophone it would be possible to relay the sound output of the optophone to a blind user without interfering with their auditory perception of the surroundings. Although useful, this device was left in the initial stages of research with respect to combining with the optophonic output, since it was believed more important to search for an optophonic system that worked sufficiently well before investigating alternative means for presenting the sound output to the user. Furthermore, the sound heard from the Neurophone is mono rather than stereo, which severely limits the ability to perceive direction with an optophone. An area of future development would be to extend the capabilities to stereo sound, possibly through the use of two separate Neurophones.

6.3. Foveated images

The human vision system has a visual field that is approximately 180 degrees wide and 120 degrees high, with a narrower central field of binocular vision. This region of high resolution (macula lutea) at the centre of gaze is roughly oval in shape, covering a field of view of about 12 degrees wide and 4 degrees high. Within the macula lutea

there is a circular region of even higher resolution (fovea centralis) occupying about one degree of view. [Dew99]

Knowing this it is possible to modify images so they consist of fewer intensity changes (regions having a lower complexity), whilst providing an observer with the same detail as obtained from an unmodified image. Foveated imaging systems accomplish this reduction in transmission bandwidth of images by attempting to map them to the spatial resolution of the human eye. These procedures are based on the fact that the spatial resolution of the human eye is space variant, decreasing with increasing eccentricity from the point of gaze.

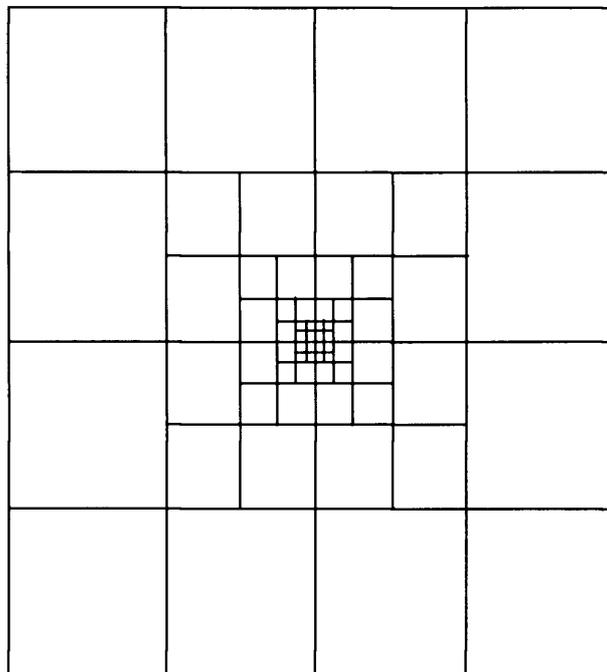


Figure (6.1) – A possible resolution setting for image foveation with central fixation. SuperPixels double in size away from the fovea fixation point.

In this way it is possible to generate an image that to the human eye is perceptually indistinguishable from a constant resolution image, but requires a fraction of the bandwidth. [Dew99, GeiPer98, GeiPer99, KorGei96, Ohe94]

Researchers, such as Geisler, Kortum, & Perry [GeiPer98, GeiPer99, KorGei96], found that by using foveated systems it was possible to process 256x256 pixel images at rates of 20-30 frames per second on standard computers, with at least a 15 fold reduction in bandwidth. The technique used employs a pyramidal approach that sequentially performs a low pass filter followed by a down sampling by a factor of 2 in all directions on duplicates of previous images. After a few iterations of this process the modified images are recombined, building a foveated image centred about the region of user interest. For instance, a window is selected at the fovea and filled with pixels from the highest resolution image (layer 1). Next, a region around the first is chosen and filled with pixels from the second slightly lower resolution image

(layer 2). This process is repeated until all layers have been combined (figures (6.1) & (6.2)). During recombination interpolation can be employed to reduce the jagged appearance apparent in figure (6.2).

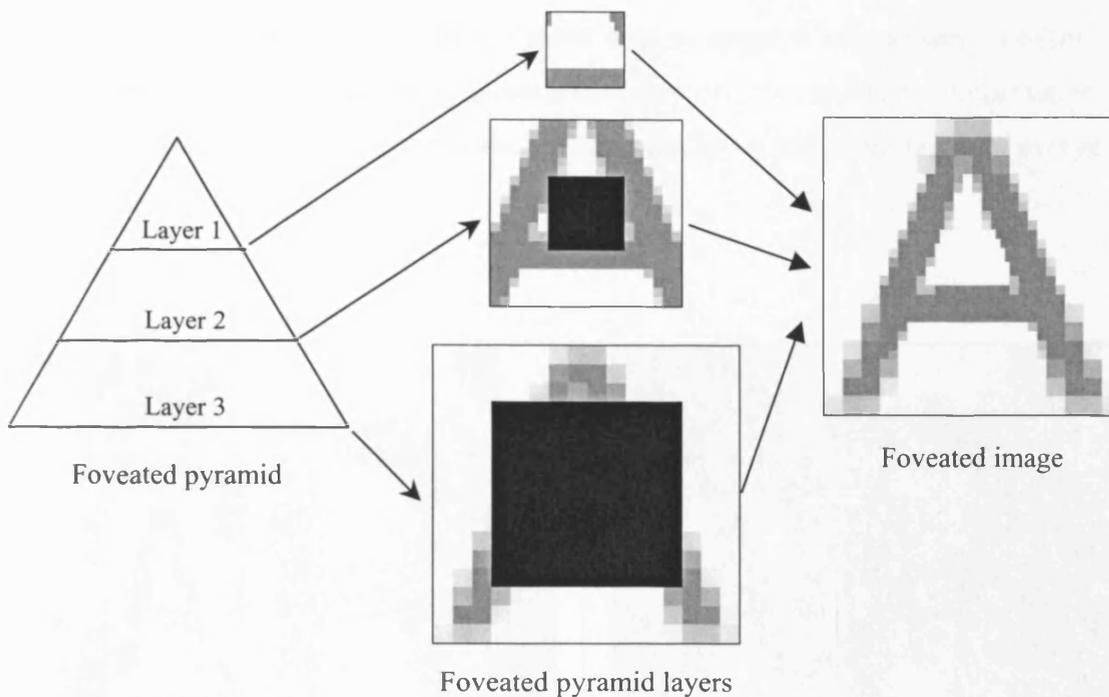


Figure (6.2) – Simplified graphical representation of the process using foveated pyramids to create foveated images.

Figures (6.3b) & (6.3d) demonstrate examples of centrally foveated images that have used four layers each with a resolution half that of the previous. The quantity of pixels in these images is 256×256 , whereas the number of SuperPixels [KorGei96] obeys a similar layout to that shown in figure (6.1) and is equal to 5824 (central region, layer 1, consisting of $1024 = 32 \times 32 \times 1$ SuperPixels, layer 2 with $3072 = 12 \times 16 \times 16$, layer 3 with $1280 = 20 \times 8 \times 8$, and layer 4, the outer ring, with $448 = 28 \times 4 \times 4$). With a SuperPixel corresponding to the size of a group of pixels determined by the layer or resolution used at a particular image location. Assuming that for figures (6.3a) & (6.3c) the size of a SuperPixel throughout the images is equal to one standard pixel, then there has been a reduction in SuperPixels by about a factor of 11 ($65536/5824$) between the foveated images and the originals.

Although this technique reduces the number of differing intensities encountered in any one region other than the fovea, consequently reducing the overall bandwidth of the images, it does cause some problems. Firstly, with the current optophonic system the whole captured image (the 64x64 pixel display) can be considered as the fovea, so further reducing the resolution of the outer regions appears unnecessary, resulting in further distortion of an already downgraded image. Secondly, no information is provided about depth within the scene, which was felt an important factor in everyday mobility.



Figure (6.3a) – Simple scene with dimensions 256x256 pixels, consisting of letter boards 'A' & 'B'.

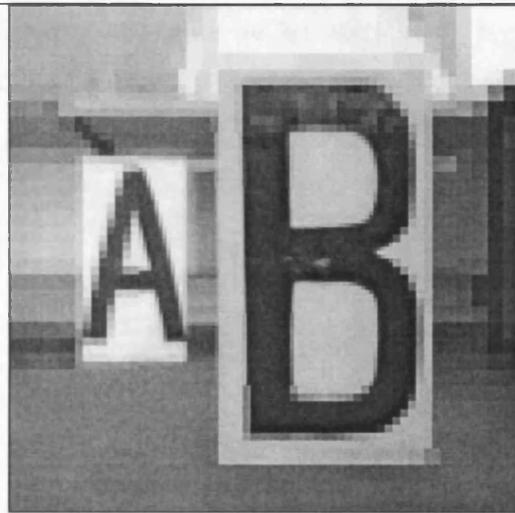


Figure (6.3b) – A 4 layer centrally foveated version of figure (6.3a) with each layer having a resolution half that of the previous.

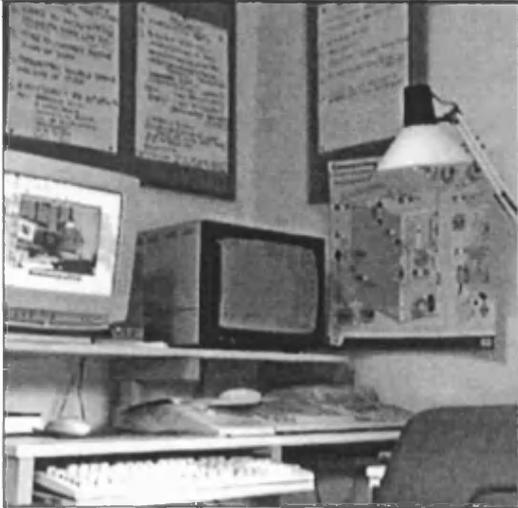


Figure (6.3c) – Office scene with dimensions 256x256 pixels.



Figure (6.3d) – A 4 layer centrally foveated version of figure (6.3c) with each layer having a resolution half that of the previous.

6.4. Monocular stereovision

6.4.1. Translational fundamental matrix

The stereo techniques described in previous sections work well with two cameras, and very well with three cameras. However, the use of a multi-camera unit is not always feasible due to equipment bulk, expense, etc. For this reason investigations were undertaken to consider methods for generating stereo depth maps from a monocular system.

Problems exist when considering monocular camera systems for generating depth maps. For instance, it is difficult to calculate range information from one image frame. Two or more frames must be taken whilst the camera is moving, preferably in a translational fashion. If this is the case, the frames can be treated as normal stereo images using methods for generating stereo depth maps. However, there still remains the problem of determining the direction of motion between frames before the depth map is generated.

The method proposed for creating edge depth maps from a sequence of images, after undergoing translational motion, is based upon the Translational Fundamental Matrix [Arm96]. Once computed, this enables the calculation of the epipolar lines for matching any features in the respective image pair. The epipolar line corresponding to the line along which features in the two stereo images lie due to camera motion. Prior to this, a system is required for detecting accurate and distinctive features in the chosen images, which can be used to find correspondences. These features are used for initial matching, and must be distinct to be readily matched, and sparse to save on computation time. These criteria generally discount edge features as a suitable first match, since any point on an edge can usually be matched with any other point on that same edge. Instead effective features are found in the form of corners, or in small regions of the image that vary rapidly in intensity in at least one direction, since these are generally fewer and farther between and comparatively easy to match.

There are many different feature detectors [GenMor76, HarSte88, JaiKasSch95, KorZim86, Mor96], but the one chosen was the Moravec Interest operator [GenMor76, JaiKasSch95, Mor96] that detects the latter type of feature described above. This operator can be used quickly and effectively to locate candidate features by taking measures of the intensity change over successive regions within the image. Peak values are then selected, and a minimum threshold imposed to give a distinct set of features.

This is achieved in the following way:

- For every pixel in the image (of a pre-smoothed image), calculate the sums of the squares of pixel differences in the four directions (horizontally, vertically, and diagonally) over an $n \times n$ window (where n is an odd number of pixels; 3, 5, 7, 9, ...).
- Take the smallest value from the four calculated and use it as a measure of interest. The larger the value, the greater the interest point. At this point a threshold can be applied.
- Suppress all non-maxima values.

Since edge points are not sufficiently distinct, whenever possible they should be avoided for the purposes of initial correlation. For this reason the system devised utilised an additional constraint to the maxima thinning routine to further reduce the number of detected features, and to improve the likelihood of making good

correspondences. To find 'strong' features, it is necessary to look for points that, when applied to the Moravec interest operator, give minimum and maximum values that are nearly identical. For example, a bold edge would result in a large difference between minimum and maximum variances, however, only distinctive image features would have nearly the same variance in all directions.

This initial stage of the process looks for a few strong correspondences to enable the calculation of the translational fundamental matrix to proceed. Consequently, the operation is made more efficient, or faster, by reducing the dimensions of the stereo images prior to feature detection. For example, if the original images have dimensions of 320x240 pixels, then good results are still attainable from a set of reduced images of size 80x60 pixels. This has an added bonus of lowering the computation time by decreasing the maximum search range, or disparity. For instance, the 320x240 pixel images shown in figures (6.4a) & (6.4b) were captured through a monochrome camera whilst undergoing a horizontal displacement. Figures (6.4c) & (6.4d) are the respective images after reduction (to 80x60 pixels), followed by a 9x9 Moravec interest operator and thinning routine.

After finding the two sets of image features, a few strong/likely matches are found by performing an initial correlation routine. Classically, correlation can be used to search an image for a known shape or template, which can be rather time consuming. This is one reason for the necessity of a small quantity of distinct features for matching; otherwise this procedure is too lengthy to successfully run in real-time.

For every feature in the first image, a small window of pixels is taken as a template and compared with the pixels located around the equivalent position in the second image. As it is rare that an exact match will be found due to inconsistencies in lighting, a measure of dissimilarity can be used to try and compensate. An effective measure that can be used is the pseudo-normalised correlation, equation [4.7] section (4.2.1), with a match being given by the correlation value closest to one.

In figures (6.4e) & (6.4h), portraying matched correspondences that indicate a specific camera displacement (horizontal), some obvious errors are found to occur. Also a number of correspondences are made that almost, but not quite fit the expected. For

example, the correspondences made are one or two pixels away from that expected for a perfect match. In figure (6.4e) two such near horizontal matches are apparent, and six lie very close to the horizontal in figure (6.4h). These are fairly reliable matches and do not adversely effect the results due to any minor deviation from the true camera displacement. These inaccuracies arise through the process of averaging whilst performing image reduction and through an additional stage of pre-filtering with a low pass Gaussian approximation. This, in combination with the interest thinning routine and the finite resolution of an image, can shift interest points, causing correspondences to sporadically drift away from the ideal.

Figure (6.4e) shows the correspondences found by applying the 9x9 pseudo-normalised correlation to figures (6.4a) & (6.4b), about the Moravec interest points denoted in figures (6.4c) & (6.4d). The correspondences themselves are portrayed with lines joining the respective image coordinates. In this example, nine apparently ideal correspondences are made out of the 12 found. Of the remaining 3 correspondences, two have near horizontal gradients and can be taken as reliable since they result from slight inaccuracies due to image averaging. Hence, 92% of all matches are reliable, indicating a horizontal camera displacement between the image frames (figures (6.4a) & (6.4b)), enabling the generation of a stereo depth map.

Similarly, figure (6.4h) represents the same procedure as described for figure (6.4e), however, rather than using image dimensions of 80x60 pixels, figure (6.4h) represents correspondences found from 160x120 pixel versions of figures (6.4a) & (6.4b). In this case, of the 29 lines shown, 14 demonstrate ideal correspondences, whilst six others are one pixel out of place due to averaging errors. Hence, 69% of all matches suggest a horizontal displacement between the stereo images, figures (6.4a) & (6.4b).

Due to the speed of operation of the algorithm, which is near real-time, image frames from a moving camera can be processed several times a second, so keeping the maximum disparity between images small. This further speeds up the process by allowing the selection of a smaller search range for correct matches, resulting in a more favourable ratio of correct to incorrect matches.

Although the use of thresholding is often frowned upon due to its tendency to limit the capabilities of the technique in question, in trials it was found that restricting the maximum disparity prior to the correlation search improves the results. If it is known that there will be a wide range of disparities encountered from a set of images, then the ratio of correct to incorrect matches will increase as the maximum allowable disparity is decreased, as demonstrated in figures (6.4e) & (6.4h). This relationship whereby an increase in accuracy is achieved by a decrease in search range is explained by considering that there is only one true match for each pixel/feature in the images. The implication of this is that if the maximum allowable disparity is increased, then there are a greater number of pixels to compare within the search range, and thus, a greater quantity of likely (but false) candidates, which has a detrimental effect on the overall performance. For this reason, when searching for correspondences it is often advantageous to perform a fast search with the maximum disparity range set to some fraction of the image width (for example, a 10th). If the number of good matches encountered are fewer than that required to deduce the translational fundamental matrix, i.e. no support is given to parallel epipolar lines or an epipolar crossing point, then the maximum search range is doubled. However, it is rare that this second search is required.

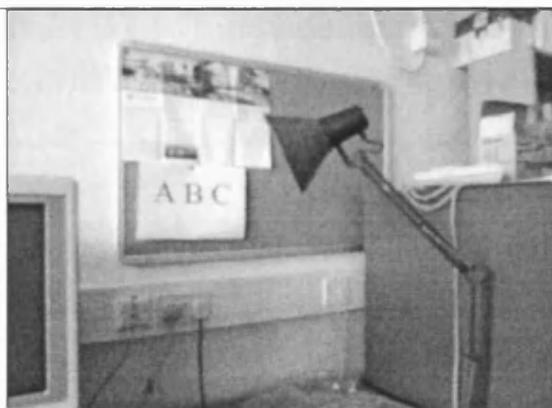


Figure (6.4a) – Image one of stereo pair. Dimensions 320x240 pixels, and a pixel intensity of 0-255.

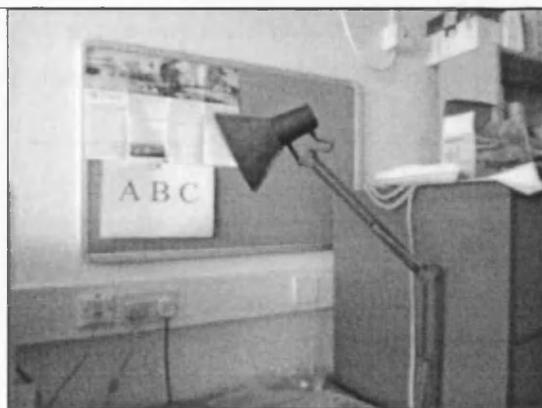


Figure (6.4b) – Image two of stereo pair.

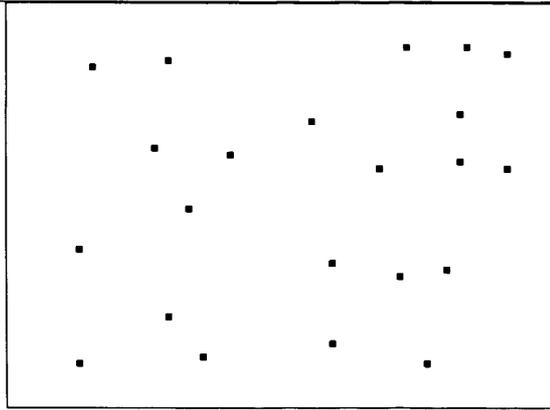
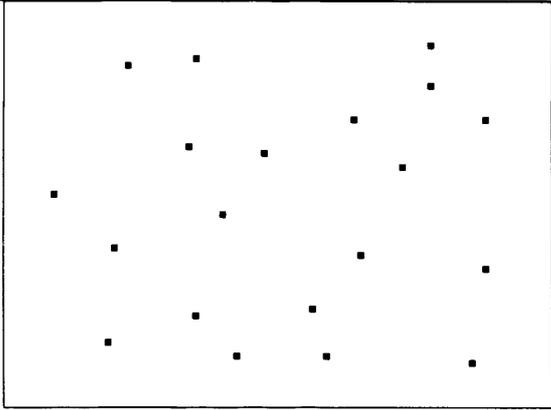


Figure (6.4c) – 9x9 Moravec interest operator and thinning routine applied to figure (6.4a) after reducing its dimensions to 80x60 pixels.

Figure (6.4d) – Moravec interest operator and thinning routine applied to figure (6.4b) after image reduction to 80x60 pixels.

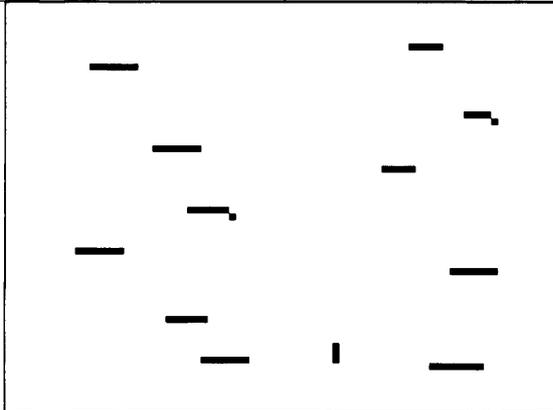
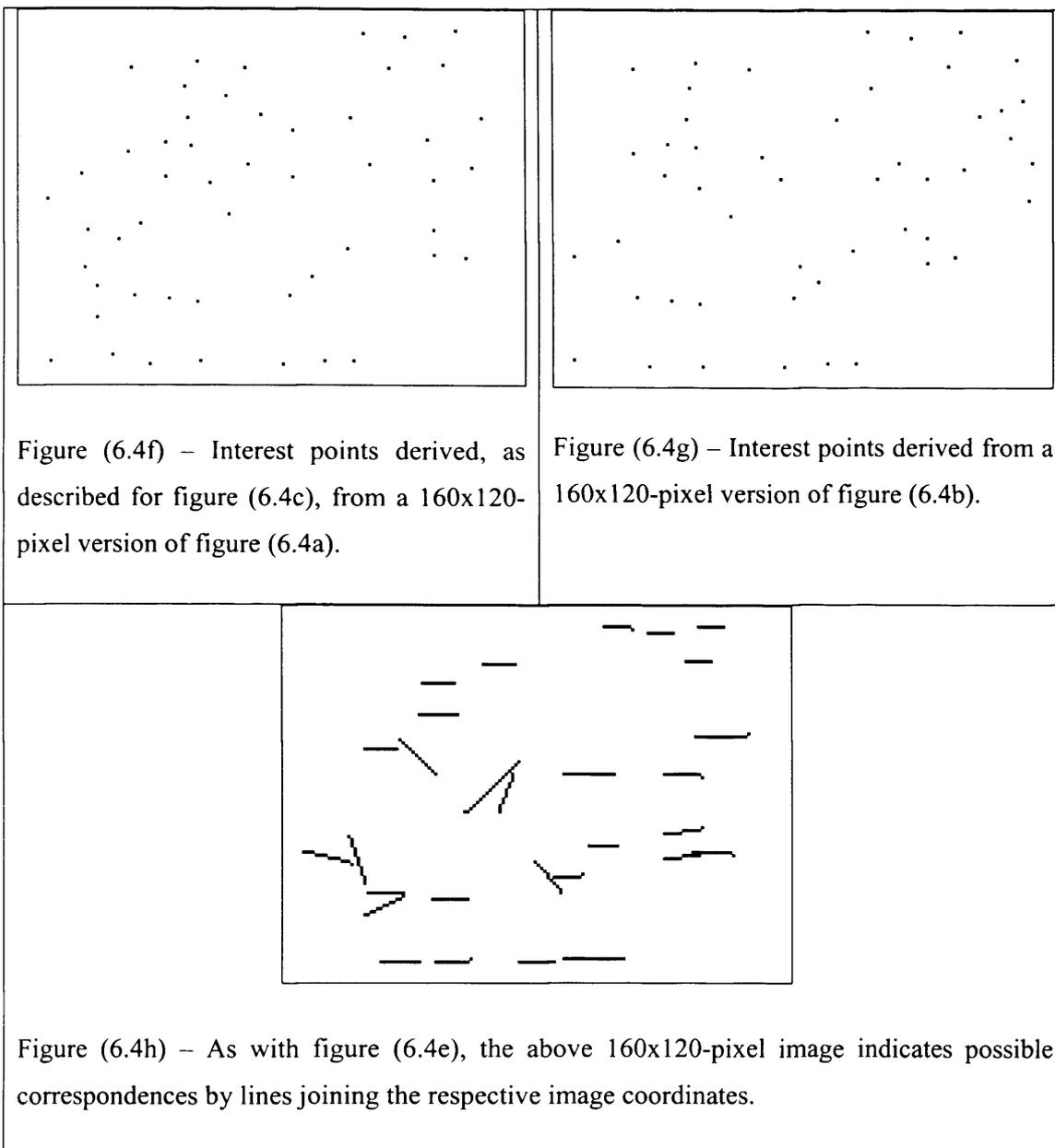


Figure (6.4e) – Correspondences found by applying pseudo-normalised correlation to figures (6.4a) & (6.4b), about interest points denoted in figures (6.4c) & (6.4d).



6.4.2. Derivation of the translational fundamental matrix

As seen from figures (6.4e) & (6.4h), if the camera motion under consideration was purely sideways, then there is little or no problem in calculating the epipolar line. Taking the median (or mean) angle or gradient of the matches would reveal a horizontal direction of optical flow, which then enables the generation of a depth map via any one of a multitude of stereo algorithms. To extend the described method to include the possibility of translational motion into and out of the scene (as of a person

walking along a corridor), the translational fundamental matrix must be calculated [Arm96] as defined by equation [6.1].

$$\mathbf{F} = \begin{pmatrix} 0 & F_{12} & F_{13} \\ -F_{12} & 0 & F_{23} \\ -F_{13} & -F_{23} & 0 \end{pmatrix} \quad [6.1]$$

Therefore, for a camera undergoing translational motion between two frames, with a set of homogenous image points $\{\underline{x}_i\}$ in the first view, which are transformed to the set of points $\{\underline{x}'_i\}$ in the second, then the 3x3 matrix, \mathbf{F} , satisfies the epipolar constraint given by equation [6.2].

$$\underline{x}'_i{}^T \mathbf{F} \underline{x}_i = 0 \quad \forall i \quad [6.2]$$

Equation [6.3] can be derived for a pair of image points (x, y) and (x', y') , obtained from the first and second images, respectively, after translational camera motion.

$$(yx' - xy')F_{12} + (x' - x)F_{13} + (y' - y)F_{23} = 0 \quad [6.3]$$

Which, assuming that $\delta x = (x' - x)$ and $\delta y = (y' - y)$ represent small changes in the x and y direction, respectively, can be rewritten as shown in equation [6.4].

$$(y\delta x - x\delta y)F_{12} + \delta x F_{13} + \delta y F_{23} = 0 \quad [6.4]$$

As the gradient of the epipolar line tends to zero, so will δy . Conversely, as the gradient tends to infinity, δx will tend to zero. Hence, the crossing point of all epipolar lines between the two images, when considering camera motion into or out of the scene, is defined as in equations [6.5].

$$\begin{aligned}
y_e &= -\frac{F_{13}}{F_{12}} \\
x_e &= \frac{F_{23}}{F_{12}}
\end{aligned}
\tag{6.5}$$

For the case shown in figures (6.5a) & (6.5b), a good estimate for the translation fundamental matrix is given by using the equation of a line to find epipolar crossing points lying within the image frame. Every time a crossing point is found within the image frame, a cell within a 2-D array is incremented. A fairly reliable estimate is then found by searching for the greatest local maxima.

To take into account situations where the crossing point lies outside the image frame a different method is used, applying a pyramidal approach that initially finds the range of all values in the x and y directions. Since it is often impractical to use large arrays on a computer, the range of coordinates is initially broken down into a smaller low-resolution array (i.e. 10x10, or 100x100 cells). The relevant cells are incremented according to the epipolar crossing points encountered, and the array scanned for the greatest local maxima. This process is repeated, each time refining the search about the region containing the local maxima, until a suitable approximation is obtained for the epipolar crossing point.

Figure (6.5e) represents correspondences found using the correlation process on two 80x60 pixel image frames (figures (6.5a) & (6.5b)) that demonstrate some form of translational motion. From this ten out of thirteen (77%) of the matches obtained indicate the existence of an epipolar crossing point within the image frame, implying the camera undertook a translational motion into or out of the scene between figures (6.5a) & (6.5b). Using the method of peak detection described, an epipolar crossing point can be obtained that corresponds to coordinates (140, 108), represented as the black dot just below and left of the crossed white lines pictured in figures (6.5a) & (6.5b).

Likewise, figure (6.5h) corresponds to matches found between two 160x120 pixel versions of figures (6.5a) & (6.5b). In this figure, approximately 40 good correspondences appear to have been made out of the 69 shown, all of which

converge left of centre of the scene. Hence, 58% of all matches obtained indicate the existence of an epipolar crossing point within the image frame, implying that the camera undertook a translational motion into or out of the scene between figures (6.5a) & (6.5b). The remaining correspondences either support the epipolar crossing point to a lesser extent, or are fairly random and tend not to support any form of acceptable epipolar lines. In this case the epipolar crossing point corresponds to coordinates (122, 116), pictured as the black dot left of the actual crossing point, represented by the intersection of the white lines in figures (6.5a) & (6.5b).

It can be seen that the smaller input images (figures (6.5c) & (6.5d) rather than figures (6.5f) & (6.5g)) generate fewer matches, but of those found a more favourable ratio of correct to incorrect correspondences is obtained.

Using either of the estimates for the epipolar crossing points (obtained from the correspondences in figure (6.5e) or (6.5h)) it is possible to calculate the translational fundamental matrix (\mathbf{F}), and generate a depth map. If it is assumed that $F_{23} = 1$, using the coordinates obtained from figure (6.5e) and equations [6.5], then a reasonable estimate for \mathbf{F} is given by equation [6.6].

$$\mathbf{F} = \begin{pmatrix} 0 & 0.00714 & 0.77143 \\ -0.00714 & 0 & 1 \\ -0.77143 & -1 & 0 \end{pmatrix} \quad [6.6]$$

Once the fundamental matrix has been calculated, it is possible to determine the epipolar lines at any point, and then a simple stereo edge detector can be applied to generate an edge depth map (figure (6.5i)), albeit slightly distorted due to the nature of translational motion into and out of the scene.

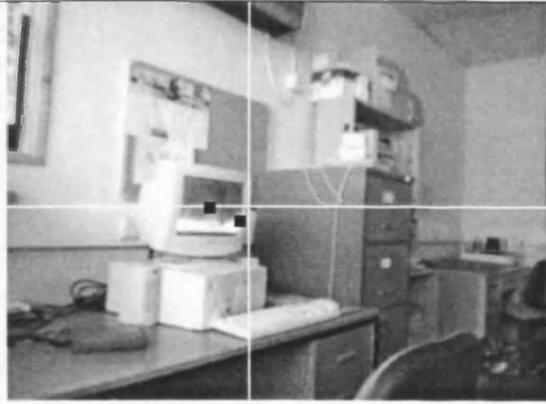


Figure (6.5a) – Frame 1 – Office scene 320x240 pixels. The crossed white lines, calculated by visual inspection and intersecting at (145, 117), represent the scene point at which the camera is withdrawing from with respect to frame 2 (figure (6.5b)).

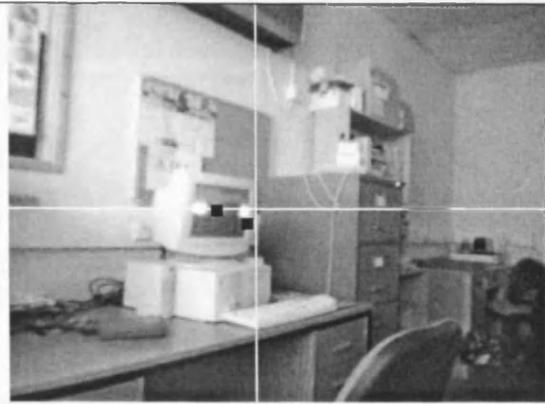


Figure (6.5b) – Frame 2 – Camera motion between this and the previous frame is away from the intersecting white lines pictured near the centre of the scene.

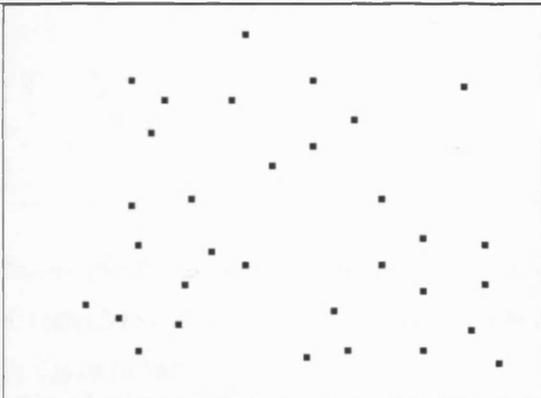


Figure (6.5c) – Interest points generated from an 80x60 pixel version of figure (6.5a).

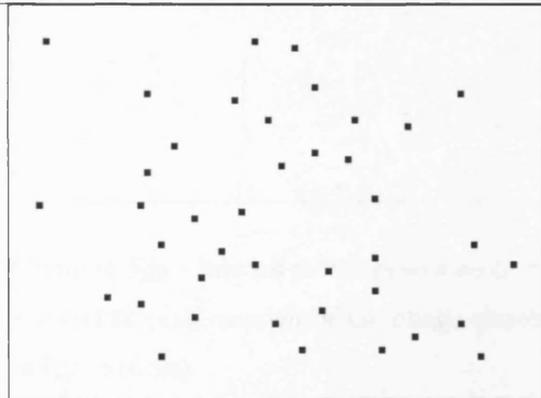


Figure (6.5d) – Interest points generated from an 80x60 pixel version of figure (6.5b).

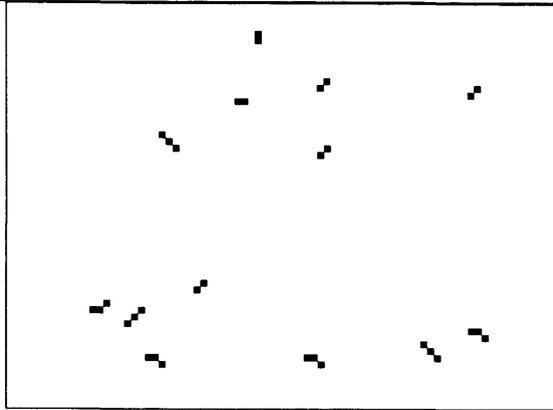


Figure (6.5e) – Correspondences found using pseudo-normalised correlation on figures (6.5a) & (6.5b) about the points shown in figures (6.5c) & (6.5d).

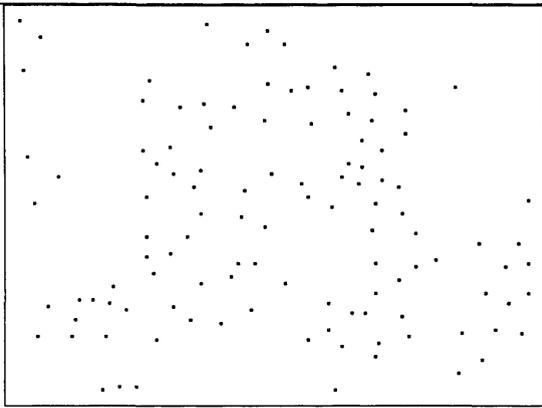
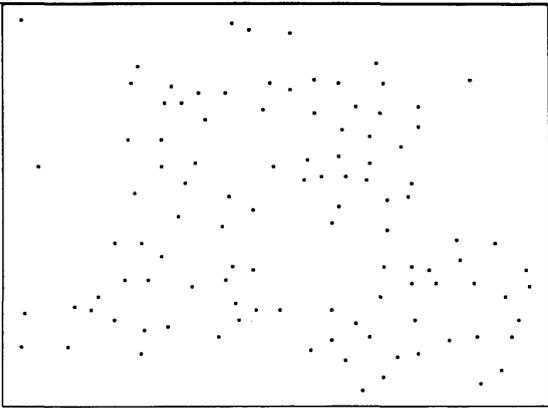


Figure (6.5f) – Interest points generated from a 160x120 pixel version of the image shown in figure (6.5a).

Figure (6.5g) – Interest points generated from a 160x120 pixel version of the image shown in figure (6.5b).

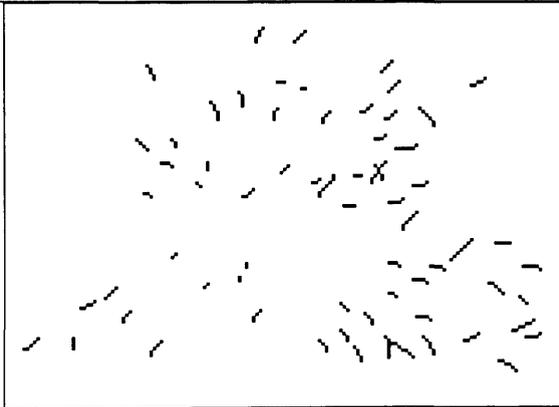


Figure (6.5h) – Correspondences found from figures (6.5a) & (6.5b), using pseudo-normalised correlation about the points shown in figures (6.5f) & (6.5g).

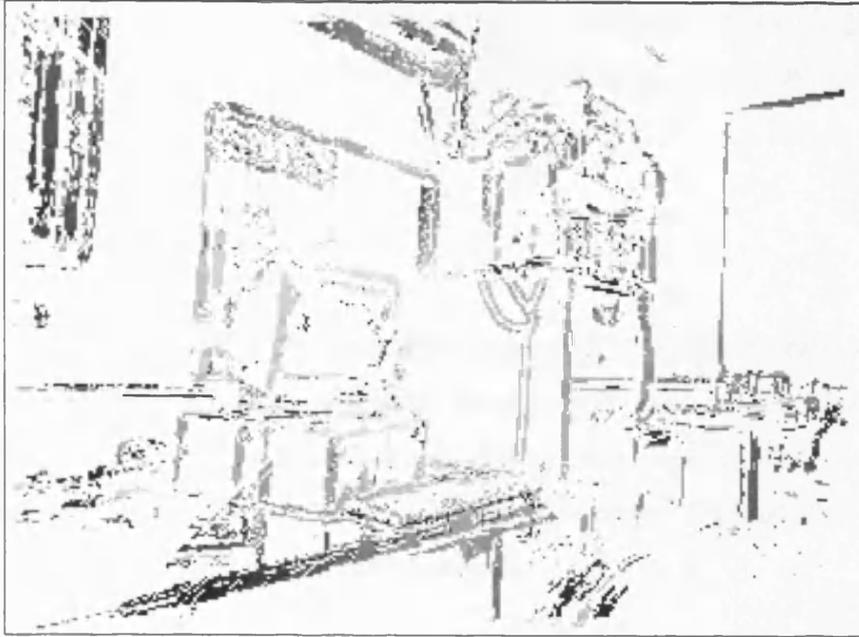


Figure (6.5i) – A simple stereo edge detector applied to figures (6.5a) & (6.5b), with the assumed intersection of all epipolar lines at coordinates (140, 108). The second intersection point derived for figure (6.5h) can also be used to generate the above image with almost no difference in result. Similarly, the true intersection (145, 117) harbours almost no further improvement.

Using the method described with images reduced from 320x240 to 80x60 pixels, an accuracy of about 75-90% was achieved in all tests, whereas with images that were only reduced to 160x120 pixels an accuracy of about 60-70% was obtained. Applying this process to the reduced images, along with a suitable stereo algorithm utilising the calculated epipolar lines, and the subsequent image-to-sound mapping, it was possible to process several frames per second. To process larger frames, or to calculate the full fundamental matrix requires a much greater length of time, consequently it was concluded that it was unsuitable for use with a software-driven optophonic blind aid.

The translational fundamental matrix, although proven to work with simple camera motion, was found to be limited in its usefulness. It is rare that a person, while walking, moves with perfect translational movement. The method described would fail, or produce spurious results, at the slightest deviation from perfect translational motion, such as the user varying their angle of movement. Furthermore, as a person

walks, their gait produces a complex motion that would hinder accurate calculation of the epipolar lines between consecutive images, thus worsening the problem.

6.4.3. The Hough solution

A slightly different approach to stereo processing that was investigated uses a line detector prior to matching [MarDurCha85, MciMut88]. Using this method it is also possible to use multiple images from a monocular camera system whilst undertaking some form of motion between frames. A method for detecting lines that is suitable for this form of matching is the Hough transform.

An edge operator that is capable of accurately providing edge orientation is used to locate edge pixels prior to placing into the Hough space. Once complete the edge pixels are mapped into the Hough domain to form lines that are categorised under their orientation in the image plane. In this technique, edge points vote (in Hough space) for all of the straight lines, to which they could belong. Line parameters that receive sufficient votes correspond to lines in the image. The two Hough domains (one for each of the stereo images) are then scanned looking for lines that have near identical orientations and length. If a match is found, then a line can be placed in the depth map that corresponds to the location and range of the matched edges.

Ideally, when complete, a depth map is formed that consists of real-world object lines of varying intensity depending on their distance from the observing cameras. Unfortunately, this is not always the case. Figure (6.6e) shows an example of a line depth map generated from a computer algorithm (albeit a very rudimentary algorithm) using the Hough technique. Many lines have been excluded from the depth map and some have been incorrectly matched. These errors often occur due to differences in the lengths of the detected lines between the two images. For instance, a line may not be matched correctly if it differs by more than a couple of pixels between the two images.

Improvements may be made to this system by enhancing the use of the Hough domain to better detect lines of edges in the stereo images. The problem with this idea is that

the depth maps (figure (6.6e)) require at least one second to compute one frame. With further enhancement of the algorithm to improve the resulting depth map, computation time will also be found to increase considerably.

Further limitations of this form of stereo system come from the fact that the Hough transform revolves around edge orientation. Within the standard Hough domain all edges are classified by their orientation in the image. If any form of rotation occurs between the stereo images, then matching lines between the two Hough spaces will have differing angles and will be categorised differently. Thus, correct correspondences become nearly impossible to make.

Even without any form of rotation between the stereo images, using the Hough domain any line is a possible candidate for any other line within the images as long as they have similar orientations. In this case it is sometimes possible for clearly incorrect matches to be made from opposite sides of the images.

McIntosh & Mutch [MciMut88] professed a solution to the problem of quality of depth map by involving a line detector that provided a greater quantity of information for each line detected. With this extra information it was possible to discount a large number of false candidate matches. The technique proposed used a method of forming straight-line segments developed by Burns, Hanson, & Riseman [BurHanRis86]. Eight descriptive features for each line were used for the purposes of matching. These features included the line end points, angle, and length, as well as the line width, contrast, steepness, and light and dark edges.

In the paper by McIntosh & Mutch [MciMut88] it is stated that straight line matching has a considerable advantage over point matching. They claimed that a line segment generally has fewer candidate matches since there are fewer lines in an image than points, and they are usually more distinctive. Secondly, due to the lesser number of lines to match, the computation time should be lower than matching points. However, during research for this thesis, whilst investigating the Hough technique for use with stereo images, the contrary was observed. It was found that the time required in generating a pair of line images prior to stereo matching far outweighed the period necessary for simple edge detection and point matching combined.

Assuming that the above problems do not occur and a correct match has been made, there still remains one further difficulty. Accurately determining the depth of a pair of matching lines obtained from a pair of images representing some form of unknown translational camera motion necessitates the identification of their true end points. However, the end points of object lines within the images are often obscured by other objects or are simply incorrectly identified due to the process of edge detection or thresholding when applying the Hough transform. Without knowing the true location of the end points of the two lines the epipolar line cannot be calculated with any certainty. The implication of this is that lines are often misrepresented in the final depth map.

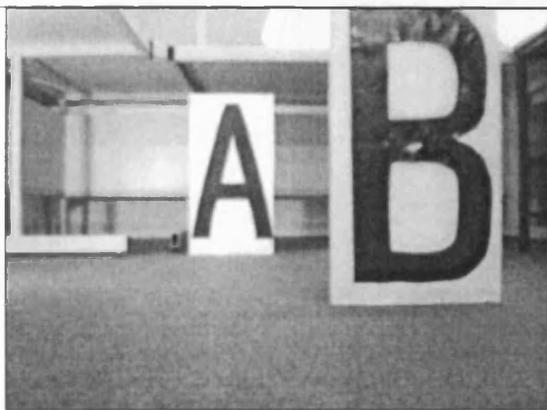


Figure (6.6a) – Left image of a stereo pair showing cardboard letters 'A' & 'B'.



Figure (6.6b) – Right image of a stereo pair.

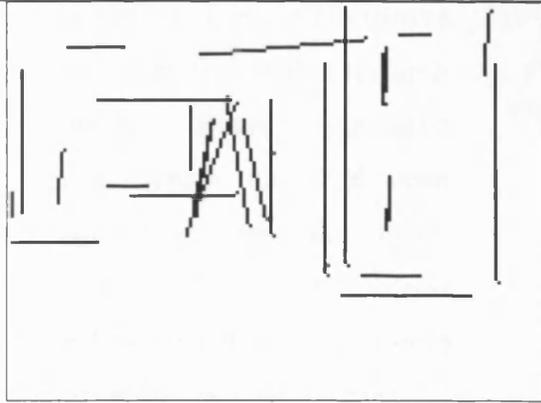


Figure (6.6c) – Hough image of figure (6.6a). Edge and peak thinning has been employed to generate a very basic edge/line image.

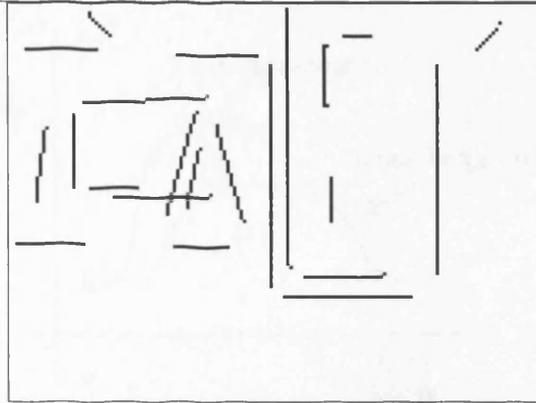


Figure (6.6d) – Hough image of figure (6.6b).

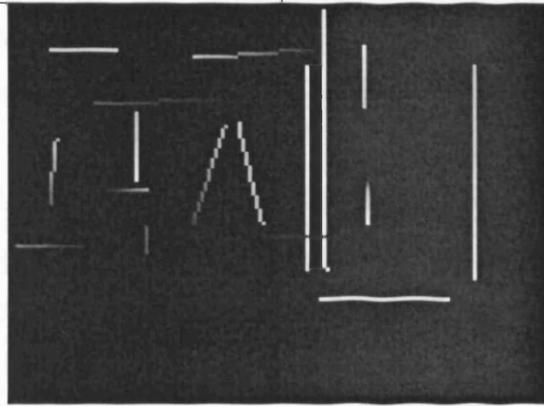


Figure (6.6e) – A very basic line depth map formed from the information provided by figures (6.6c) & (6.6d). Distance is indicated by the intensity of the lines – light is near, dark is far.

6.4.4. Depth from defocus

As mentioned in section (4.2), there is another non-standard method for determining the depth of objects within a real-world scene. The process uses a selection of images of a scene taken from a fixed camera. Each captured image should be taken with the lens of the camera focused at a different range. By applying a series of computationally intensive calculations it is possible to obtain an estimate for the depth of different regions within the scene. [GeiKla94, KlaGeiBov95, RajCha97]

A quicker, more basic alternative that was attempted in this research was to use a single image taken from a camera that had been focused on, or about, the closest object in the scene. All objects viewed beyond this range would appear with varying degrees of blurring depending on their distance relative to the observing

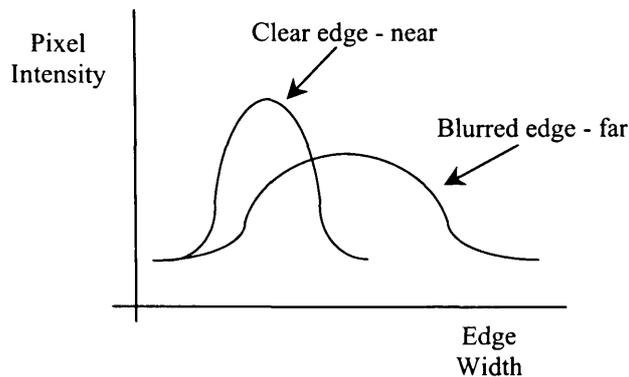


Figure (6.7) – Simplified representation of the effect that blurring has on image edges.

camera. The more out of focus the image region, the less high frequency image detail that is present. Using this principle a high-pass filter (edge filter) could be applied to provide some indication as to the distance of object edges within the image, such that the edge width increased with range (blurring), as demonstrated in figure (6.7).

The detection of wide edges is not difficult, for example, one solution is to simply apply a larger edge operator. This means that the larger the edge filter applied to an unfocused image, the more edges detected. By applying a range of varying sized Laplacian filters to the unfocused image, with each edge map corresponding to a different range or distance, a depth map is created. For instance, if three different filter sizes were used, the generated edge maps would be encoded so that the first consisted of black edges, the second with grey edges, and the last with light grey edges (background of white). The depth map would be formed by a combination of the three edge images, in each case taking the darkest pixel. The reason for this method of combination is explained by remembering that a large edge filter detects both clear and blurred edges, whereas a small edge filter only detects the thinner, clearer edges (figure (6.8a)-(6.8d)). If at the same location within an image both dark and light pixels are detected with a small and large filter, respectively, then an object exists close to the camera in the equivalent position in the real world. However, if only a light pixel is found, then this implies the object is further away from the observing camera, and is undetectable by a smaller edge filter because it is blurred to a greater extent.

The different edges (due to blurring) that are detected by edge filters of differing sizes are apparent in figures (6.8b) & (6.8c). In particular the top two photos of figure (6.8c) reveal several regions of background detail (plants) that are blurred beyond the point detectable by the 5x5 filter used to generate figure (6.8b), indicating the greater distance of these features within the scene.

The depth map shown in figure (6.8d) consists of a combination of five images that were created with edge filter sizes of 3x3, 5x5, 7x7, 9x9, & 11x11, such as in figures (6.8b) & (6.8c). After, or during edge detection, all edge pixels are converted to a predefined intensity, and all blank areas are set to white (intensity of 255). The edge intensity is determined by the number of filter sizes being used and whether a linear scale, or otherwise, is going to be used. In the case of figure (6.8d), the pixel intensities were determined as shown in Table (6.1).

A = Index (filter size)	1 (3x3)	2 (5x5)	3 (7x7)	4 (9x9)	5 (11x11)
B = A*A	1	4	9	16	25
	Find pixel intensity multiplier : $10.2 = 255 / 25 = \text{max intensity}/\text{max B}$				
C = B*multiplier	10	41	92	163	255
Pixel Intensity = C-multiplier	0	31	82	153	245

Table (6.1) – Pixel intensities for figure (6.8d).



Figure (6.8a) – Three photos all of which are focused on the foreground, leaving the background out of focus (blurred).

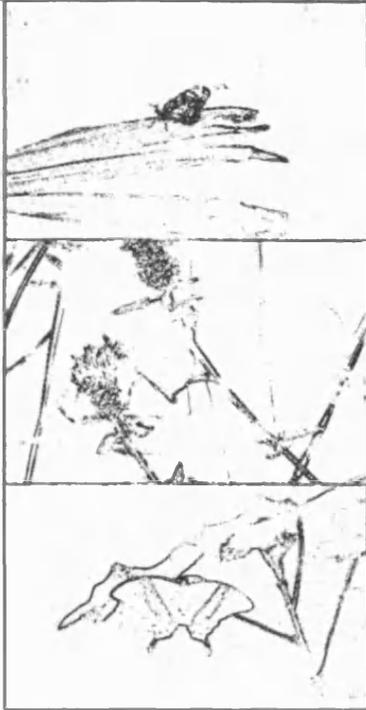


Figure (6.8b) – The result of applying a 5x5 binary Laplacian filter to figure (6.8a). All edges above threshold were set to a pixel intensity of 31 (out of 0-255).



Figure (6.8c) – The result of applying a 9x9 binary Laplacian filter to figure (6.8a). All edges above threshold were set to a pixel intensity of 153 (out of 0-255).



Figure (6.8d) – The depth map formed by the combination of figures (6.8b) & (6.8c), as well as three others that correspond to filter sizes of 3x3, 7x7, & 11x11, and pixel intensities of 0, 82, & 245 respectively (0 corresponds to black).

Applying a series of increasingly sized edge operators to an image is rather time consuming, to the extent that only 1 or 2 frames per second are processed for reasonably high-resolution images (about 640x480). However, a pyramidal approach [Fua93, JaiKasSch95] can be used to decrease processing time, by reducing the image size at each stage, whilst retaining a constant sized edge filter.

The problem with this form of edge depth map is its accuracy, which is often flawed, to the extent that it may become unreliable.

6.5. Summary

Preliminary research studied the methods by which images could be modified to better mimic the workings of the human vision system. In this way it was envisioned

that a blind user could adapt more readily to the generated sound that corresponded to the modified images.

The first step towards modelling images on the human vision system was to use foveated images, which used a central high-resolution area (fovea). The images were coded so that their resolution approximately matched the spatial resolution of the human eye, decreasing in quality towards the periphery. Ideally, images created using this method would be perceptually indistinguishable from the original unaltered image consisting of a constant resolution.

The advantage of foveated images are the massive reductions in bandwidth that result from the downgrade in peripheral quality, thus reducing the amount of image data portrayed in the optophonic sound output. Nonetheless a number of disadvantages of this form of image became apparent. Firstly, there was a lack of any depth information that is an essential part of the human visual system that enables competent mobility, and secondly, the reduced image (64x64 pixels) already used with the optophonic system is, for all intense and purposes, a fovea. The angle commonly displayed by the optophone (although partially dependent on the video input device) is relatively small in comparison to the visual system's view field, and is in a sense already comparable to the fovea encountered in human vision.

Whilst investigating stereo depth maps some initial work was undertaken studying aspects of the fundamental matrix and its use in generating stereo depth maps. The fundamental matrix can be calculated from a set of matching points taken from a pair of images of a particular scene, assuming some form of camera displacement between frames. Once calculated the fundamental matrix allows the determination of the epipolar lines, which connect matching points between the image pair, for any point within the two images. With this, a depth map, albeit sometimes slightly distorted, can be generated.

The preliminary work into this area studied the translational fundamental matrix for use with a monocular system, assuming an unknown camera motion between frames in the x-, y-, or z-planes. Although the translational fundamental matrix has the limitation of only allowing translational motion between image frames, it does have

the advantage of speed over the full fundamental matrix due to the requirement of only three values to complete the matrix. To determine the full matrix, due to the quantity of unknown values, requires a great deal of processing time. Consequently, research into this area ceased in favour of a more rapid solution.

Further work into the use of monocular camera systems investigated the Hough transform as a method of storing information pertaining to the location and orientation of edges within the image scenes. From this, possible matches could be obtained by searching in Hough space for like edge orientation. Difficulties with this method present themselves when considering camera motion between frames in any form other than translational. In such an event, information upon the edge direction becomes ineffectual for detecting stereo matches, and consequently almost any edge is a possible candidate for any other.

Methods were considered that enabled the generation of low quality monocular depth maps through the use of a fixed near-focus camera. Images captured from such a device would appear sharp in the foreground, but progressively indistinct with increasing distance. This variation with range provided several possibilities for determining the depth of objects within the scene. For instance, the use of successively larger edge filters provided an indication as to the depth of object edges within the scene, albeit rather limited in quality and accuracy.

Unfortunately, with this method of depth from lack of focus, or defocus, it tends to be too unreliable in terms of accuracy, and can often require fairly long periods of time to process the image with a series of progressively larger filters.

7. Conclusions

During the course of the research into modifications to the optophonic system a number of alterations were developed and then tested, as well as numerous other techniques that were investigated to varying degrees. As might have been expected a complete solution to the blind mobility and reading problem was not forthcoming, however a few of the techniques investigated did improve some aspects of the task. An analysis was performed on the data received from a number of tests, in each case supporting the already presumed improvements.

A number of experiments were carried out in an attempt to test various theories about certain features of the optophonic process. These are described along with the corresponding areas for future development.

7.1. In the beginning...

From the start it would have been nice to know that whilst researching various aids for the blind, a solution to the general blind problem would be forthcoming, however that is not realistic, and certainly not within the scope of the research. The work carried out investigating the optophone was intended to help in terms of progressing one area of work related to aids for the blind. It should be realised from the numerous previous attempts that if any possible solution were to work it would entail a great deal of time and effort from both the researcher and the intended user. For this reason, from the outset the goal of this research was not to create the 'perfect blind solution', but rather to make improvements to an existing system – namely the video optophone.

7.2. The approach

It was decided from the outset that the best approach was to first investigate previous attempts at optophone-like devices. The main aim of this look at the past being to gather information as to why previous devices failed. It seemed reasonable to assume that these past devices had failed, otherwise why were they not in common use today.

Consequently, it was realised that the majority of past blind aids were unsuccessful due to the quantity of visual information they relayed to the user in the form of tactile or auditory signals. Many devices did not take account of the bandwidth reduction required when converting between these forms of medium.

It was decided that the next step should be to note the failings of other blind aids by studying various forms of data reduction and methods for re-stressing features within the captured images, followed by their application to an optophonic mapping from scene-to-sound. The first decision upon this task was to concentrate on the image side of the optophonic process since that was considered the phase that could best be reduced. Thus, only minor changes were made to the sound output, such as the work into stereo positioning.

Work into various image-processing techniques, with emphasis on methods of data reduction and modification, were carried out and incorporated into software versions of the video optophone. It became apparent that a sighted person should first be used to assess the effectiveness of any image processing performed during the optophonic mapping. The idea being, that if a sighted person could not successfully navigate using the modified images, then a visually impaired person could not be expected to correctly interpret the equivalent optophonic sound.

After the application of various image-processing techniques a belief of Adrian O'Hea's was re-examined and further developed. He believed that the blind should be provided with the same type of information as the human visual system is designed to supply. One aspect that had not been considered in a video optophone was that of depth. Not only could range information be used to improve the emphasis on

important features within an image, lessening the burden on the listener, it could also provide much needed knowledge about the distance to objects within their vicinity.

Stereovision systems were developed and incorporated into the already written video optophone software for testing. Tests were devised to assess the efficiency of any form of stereo algorithm utilised and applied to the final optophone system.

After the creation of a new stereo processing technique as well as a new form of depth map known as stereo cartooning, a method of categorising the process of testing the optophone system was devised and labelled DeLIA (for obstacle Detection, Location, Identification & Avoidance/Action).

Using DeLIA a series of tests were formed to assess the important aspects of blind mobility. Results obtained from preliminary tests with the optophonic stereo techniques showed great promise, and analysis of further data revealed improvements with many of the facets of DeLIA. Initial experiments were designed to test the Detection, Location, and Identification of objects within a visual scene (leaving the Avoidance stage for later evaluation).

7.3. Developments

Of the techniques tested, one that showed significant benefits was that of stereovision, which had been largely untested with an optophonic system. Software was designed and written to work in conjunction with a three-camera capturing device that provided a suitable basis to test future developments and to perform experimentation with volunteers.

Whilst studying the various forms of stereo depth map and algorithms two new methods for representing a depth map were attempted and found to work, in the sense that they could both portray the depth to an object whilst successfully lessening the burden on the listener by fading out less important features from within the scene image. One of these techniques, described in section (4.3.4), was a completely new

form of edge depth map, and although basic in concept and similarly in display, could be generated at extreme frame rates. However, this algorithm was found to produce depth maps, which albeit accurate and recognisable, were edge based in nature making it on occasion difficult even by sight to comprehend the scene. Since this difficulty made the recognition of a scene often impossible by a sighted person it was presumed too difficult to be correctly perceived by a user via sound. This is not meant as a reflection upon the listener, but rather an acknowledgement that after the conversion to sound even a simple scene can become taxing to decipher.

The second technique mentioned above, and described in section (4.3.5), gained the label of a 'cartoon depth map' or 'stereo cartooning' since it was the combination of a method known as cartooning and any suitable stereo edge algorithm available. The stereo edge algorithm used for the purposes of testing, and described in section (4.3.3), was chosen after evaluating the algorithm's results via both visual inspection and with a technique developed to provide a ratio value of performance, detailed in section (5.1.1), later found to be similar to a method known as the FPF (False Positive Fraction).

The advantages of stereo cartooning are only realised when considering its application and incorporation into the optophonic process. As with other depth maps, range is portrayed by the intensity of the image pixels, and similarly by the sound amplitude generated by the optophonic conversion. Unlike either an intensity or edge depth map the quantity of filling can be controlled by an adaptive threshold system, whereby the user selects a percentage threshold, which the algorithm uses in combination with an image histogram to fill the relative proportion of the depth image. Consequently, the user can select what proportion of the scene to leave unfilled, creating anything from a pure edge depth map to one that closely approximates an intensity depth map.

After completing the final stages of the stereo cartooning technique a method of evaluating subject performance with the modified stereo optophone was required. By considering the different aspects of general mobility and response, such as the almost completely innate reaction (identification followed by some form of response) to an obstacle in a subject's path, a method of categorising the process was formed (labelled DeLIA). This system, which as previously mentioned, pigeonholes the mobility

problem (or rather the problem of avoiding obstructions) into four classes, which are the Detection, Location, Identification, and Action or Avoidance of obstacles, provided the basis for further experimentation and evaluation of the optophonic modifications.

7.4. Experimentation & analysis

Suitable modifications to the optophonic process were developed and implemented (such as stereo depth maps), along with a method other than visual inspection for evaluating their accuracy and efficiency with respect to error, and an appropriate way in which to monitor or categorise a subject's performance with the subsequent optophonic output (using DeLIA).

7.4.1. Initial experimentation

Before progressing with more rigorous tests it was decided to begin with two rather subjective experiments on three volunteer subjects. These provided an insight into various features of the optophonic process, with emphasis on the maximum comprehensible frame rate and the user's ability to resolve the vertical and horizontal position of an object.

7.4.1.1. Test 1 – Frame rate

The first test was designed to assess a subject's performance and reaction when presented with a high frame rate optophonic sound. It had previously been noted that the maximum speed at which a user could be expected to successfully identify features from within an optophonic image (such as the test images shown in figures (5.7a)-(5.7p), pages 138-140) was approximately four frames per second. If the presented scene consisted of only two small moveable dots, then the frame rate could be increased to over 16 frames per second.

An experiment was created in which two dots were randomly placed upon a virtual scene (black dots on a white background), whereby the user was asked, by use of controls on the keyboard, to move one of the dots onto the second by listening to the sound alone. This test was used to provide information as to the efficiency of the sound display after one or two subjects commented on difficulties that they had encountered whilst attempting to judge vertical placement during earlier trials. Although no numerical data was offered by this test, it did provide a suitable indication as to the validity of the previous comment.

The experiment showed that at these high frame rates subjects were able to quickly and correctly position the second dot over the first in terms of horizontal location with almost no error. However, when it came to the vertical position (signified by the frequency) most subjects had great difficulty and often could not obtain the correct location, unless they could successfully gauge vertical distance by sweeping through the full range of frequencies (generally taking a great deal of time).

Although the error in the vertical positioning tended to be fairly small, it still suggests that using the frequency alone to provide information upon height within the real world is not enough. Alternatively, the range of frequencies used during the optophonic conversion from scene-to-sound may be better adjusted to provide a greater definition.

7.4.1.2. Test 2 – Obstacle avoidance

To test the usefulness of stereo depth maps a second test was performed to study the last aspect of DeLIA (obstacle avoidance). Using different optophonic modifications (stereo image algorithms) subjects were asked to manoeuvre through a room that contained a number of items of furniture. The original test consisted of an arrangement of chairs and tables the user was required to navigate through, using only the sound output from the optophone.

This experiment was designed to provide an insight into whether the proposed optophonic modifications could be used to aid a user in the recognition and avoidance of everyday items of furniture. Four different types of optophonic system were tested, namely the plain unmodified images, an edge depth map, the Triclops systems innate depth map, and the cartoon depth map. Of these mappings, subjects found that the plain images and the edge depth map provided almost no additional help. The reasons for this apparent lack of assistance can be placed upon the similarity in colour of the tables, chairs, and the carpet of the room used, all of which were shades of brown. This hindered recognition when using the plain images. The edge depth map, although providing accurate depth information, supplied insufficient image structure to correctly determine the location of an object. The edge information enabled the user to find object boundaries, but without touching the object it was not possible to decide on which side of the detected edge the object lay.

Using the cartoon depth map as input to the optophonic mapping subjects were able to quickly and accurately locate the gaps between the different obstacles. Although the adaptive thresholding worked well in most situations, problems arose when the subject transferred from looking at a cluttered scene to one consisting of large areas of limited texture, such as the plain carpet. In this case the adaptive thresholding went from displaying a well-presented cartoon depth map to one where, although having the correct depth, unremarkable regions were filled giving the illusion that an object lay in the user's path. Hence, the subject could be confident about their route between two objects, and then suddenly find a nonexistent obstacle appearing directly ahead.

Work is continuing in an attempt to find a more suitable alternative for the adaptive thresholding used in the stereo cartooning to remedy this problem.

Of the entire range of stereo imaging techniques incorporated into the optophonic process, which also employed the various stereo sound effects discussed in earlier chapters, the Triclops intensity depth map appeared to be the most useful for this form of experiment. By selecting an appropriate disparity subjects rapidly located the different objects in their local environment and found safe routes between them.

7.4.2. Further assessment

The next step was to devise experiments to use in conjunction with DeLIA to provide numerical data, and consequently evidence either for or against the proposed stereo techniques (stereo cartooning) when compared to the standard video optophone using unmodified images.

7.4.2.1. Test 3 – Preliminary optophonic evaluation

This test, shown in section (5.2.2.1), was designed to assess and compare subjects' performance between a chosen optophonic modification (in this case – stereo cartooning) and plain unmodified images with respect to the first three stages of DeLIA (that of Detection, Location, and Identification). The test achieved this by using a series of images captured with a three-camera stereo system, portraying two large letter boards with the characters 'A' and 'B'. These boards were positioned in different locations within the scene. Subjects were asked to identify the order of the characters in the scene, and the distance from the observing cameras. Also recorded during the tests was the number of repetitions required by each subject before they believed they had correctly identified an optophonic sound.

Using the '*paired two sample for means t-test*' with the data recorded (shown in Appendix A2 – section (A2.1)) from the experiments it was found that the method of stereo cartooning provided a significant improvement in all but one aspect. Nearly all subjects found the depth maps easier to interpret than the plain images, requiring fewer repetitions and demonstrating a higher performance in terms of correct responses for both the character depth and consequently the whole scene. The only category that showed no significant increase in accuracy was the identification of characters.

It is interesting to note that three of the best overall results (in terms of images correctly identified) were obtained from subjects with whom it could be said are musically minded, since they are all musicians. The results of which can be seen in Appendix 2 – section (A2.1), subjects 2, 4, & 7. This is an area to be studied further

to see whether certain people are naturally better at perceiving optophonic sounds than others.

7.4.2.2. Test 4 – Further optophonic testing

The second experiment (section (5.2.2.2)), as with the first, was designed to test the effectiveness of various optophonic modifications over the original video optophone. The modification again being the use of stereo cartooning prior to applying the optophonic scene-to-sound mapping.

An automated programme was written to test and record a subject's ability to recognise an optophonic sound, which corresponded to one of a set of known images, after only two minutes of practice with the test images/sounds. Half of all subjects tested were first presented with only cartoon depth images, followed by a second test with only plain images, whereas the second half were assessed in reverse order with the plain images first. The different order was used to prevent any bias towards one particular form of image.

Subjects were given the same instructions on the basic principles involved with both the video optophone and the test itself, and were then left, after the two-minute practice, to complete the test. Results obtained during the test can be found in Appendix A2 – section (A2.2). The experiment was performed on a larger group of subjects after performing sample size calculations on the preliminary results (section (5.2.2.2), pp. 120-121), which showed that a group size of twenty was sufficient to test for significance between test data sets at a 95% confidence level.

Performing the '*paired two sample for means t-test*' on the data gathered, as for the previous experiment, demonstrated that there was a significant improvement with respect to three of the four aspects. Using the cartoon depth images subjects required fewer repetitions to complete the tests, as well as exhibiting a greater accuracy with judgement of depth and consequently the total number of images, when compared to subject performance with the plain images.

The only aspect that showed no significant difference was the subjects' ability to correctly determine the characters (left-right location). Actually, although the *t-test* revealed no difference, by taking the *mean* values it was apparent that character recognition was sometimes slightly easier with the plain images, the average showing a difference of approximately one in sixteen. The range of intensities, and thus sound amplitudes, is one explanation for this. With the cartoon depth maps, if one character was near in relation to the camera, while the other was far, then the subject's attention often appeared to be fixated on the louder (closer) of the two. Consequently subjects could easily recognise the depth, but sometimes had difficulty in identifying the single (most noticeable) character. Whereas, with the equivalent scene in the form of a plain image, although one character (the distant one) would be smaller, both letters would be heard with similar amplitude, thus providing the subject with two chances to recognise the characters.

The above problem of identifying the characters did not appear to occur with the first experiment. Although the *t-test* still showed no significant difference between the character recognition, the *mean* values favoured the cartoon depth map. This can be explained by the length of the training period. With only two minutes practice the subjects had a greater difficulty recognising the order of characters from hearing only one significant letter (as in the depth map problem described above). However, with ten minutes practice there was a smaller difference between the rates of recognition, indicating that subjects were more adept at identifying characters.

7.5. Future development

The video optophone may not be perfect with respect to enabling a blind user to 'see' their surroundings. However, by applying the techniques described within this thesis to a video optophone it was shown possible to locate and sometimes recognise features and obstacles. Further more, a number of additional areas have been described during the course of the research that could be employed to increase the usability of the optophonic device, as well as mentioning several aspects that still

have room for improvement. Chapter (6) details a number of these areas of interest. As to the future development and investigation of these topics, some ideas are described below.

Neurophone

A further examination of the Neurophone (and its ability to pass sound to a listener by means other than the normal auditory channel, the ears) to assess the capabilities, bandwidth, and perceptible frequency range when using the device. If the sound output from an optophonic device could fully be transferred to the Neurophone, then a blind user would not effectively lose their sense of sound and so should still be able to hear auditory clues as to the various moving objects within their local environment (i.e., a car, or person).

Stereo cartooning

The adaptive thresholding used with the stereo cartooning technique failed to produce an adequate depth map in some circumstances where little or no texture existed within the captured scene. If a more suitable method of thresholding could be employed in those situations the cartoon depth map would provide an ideal input to the optophonic mapping.

Vertical coding

Study the use of either a different or additional property to encode height information, or examine the range of frequencies used in an attempt to provide a better spatial definition.

3-D soundscapes

If the output of the optophonic process could be encoded to form a three-dimensional soundscape, then assuming the application of a stereo depth map, any object within the viewed scene would appear to generate a sound originating from its location.

Optophonic applications In the past the optophone has generally only been considered as a blind aid. However, there may be a number of non-blind applications that could benefit through the use of such a device. Such an application could consist of situations where a user might need to keep track of a moving object as well as a number of instruments. In this case the optophone could easily provide information about the location of the moving object, whilst the subject visually studied the instruments or controls. For instance, a pilot could keep track of another unseen plane when in close proximity through the use of a 3-D optophonic system (working much like a radar, scanning 360 degrees).

Optophonic input Examples of images used as an input consist of plain unmodified images, edge images, foveated images, as well as various form of stereo depth maps. Of these the best form for reading appears to be an edge image, and similarly, the best for navigation seems to be intensity based depth maps. Thus, a further study of systems that combine various modes of view, either by user control, or by overlapping images, such as an edge image placed over an intensity depth map.

7.6. Summary

In final conclusion the techniques described in this thesis have been shown to significantly aid in the identification and comprehension of range information within a real world scene. The use of stereo imaging has also been found to reduce the stress encountered by the user when listening to the resulting optophonic sounds, which was commented on by a number of test subjects and demonstrated by the reduced repetition rate in the experiments with the stereo cartooning technique. Although it

was noticed that the stereo depth maps did seem to slightly reduce the recognition rate of characters and other such textural detail. This reduction appears to be related to the amount of training the user receives, with greater practice providing almost no difference in character recognition rate between the unmodified and depth images.

Consequently, the results indicate that a video optophone would benefit through the use of a stereo depth map as input for general mobility, and either a plain unmodified image or an edge image (or edge depth map) when trying to determine text or similar detail.

For many years people have tried to develop systems to 'solve' the blind mobility and reading problem, and although a solution may still not be forthcoming, it is hoped that the work detailed herein may in some way help future research towards those ends.

Appendix 1. Timeline of devices for the blind

The following table lists some of the major developments, in terms of inventions and systems for the blind, which have occurred over the last two centuries. The *name* corresponds to the inventors and/or association (company) behind the device described. The *year* shown corresponds to the time at which it was believed that the system being described was first revealed according to the sources available. The column labelled *device* corresponds to the name of the system being described, and the *use*, which uses the keywords shown below, represents what the system developed was intended for. Finally a brief description of the system is given and any other relevant information.

It should be noted that the information contained within the table is only as reliable as the papers used, and thus the reader must bear in mind that a great number of these devices appear to have been long forgotten and so were found only in very early papers that have been found on numerous occasions to contradict each other.

Keywords:

Mobility	Generates a 2-D sound or tactile map of a simplified image of the users view.
Obstacle	Detects obstacles in the users path – clear path indicators.
Reading	Reading aid for the blind.
Writing	Writing aid for the blind.

If the year is followed by either a '<' or '>' then the device was invented or released to the public after or before the year specified.

If the year precedes '<>' then the device was invented or made public around the year specified.

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
		Moonwriter	Writing	Allows a blind user to type and print a page of Moon text. [Moo86]
		TAM	Obstacle	A device that is worn like a watch, with a thin wire connecting to a small device/box that clips onto your jumper. When a car approaches, the noise level increases, and so the watchband starts to vibrate. (See web page www.s55wilma.demon.co.uk/info-db.html , accessed May 1999)
Braille, Louis	1824	Braille System	Reading	Uses a process of six dots to represent 63 different characters and symbols. Grade 2 Braille uses a type of shorthand, allowing reading speeds between 100-200 words per minute. Requires substantial sense of touch in the users fingers. An alternative to Braille is the Moon system, which tends to be easier to feel and read. (See web page http://www.s55wilma.demon.co.uk/braille.html accessed May 1999). In 1819 Louis Braille learnt of a system called 'night writing', invented by Captain Charles Barbier earlier in the same year. The system used 12 dots to allow nighttime battlefield communications. The Braille system was first published in 1929. (Information widely available from the internet) [Dew99, Moo86]
Moon, William	1845	Moon System of Embossed Reading	Reading	The system consists of nine (originally 9 – later 14 characters, including sequences of dots) curves and lines, which have different meanings depending on which way round they are presented. The majority of the characters in the Moon alphabet, in some way represent the letters of the normal English alphabet. Requires less feeling in a user's finger than Braille, but requires more space (book). Reading occurs from left to right, then right to left, and so forth, so that reading is continuous. The blind user need not remove their fingers from the page to locate the start of a new line. (Information widely available from the internet) [Moo86]
Hall, Frank H.	1892	Hall Braille Writer	Writing	The device is like a typewriter, but with only six keys that make up the dots in the Braille cell.
Noiszewski	1897	Elektroftalm	Obstacle/Mobility	Possibly the first mobility aid for the blind. The device uses the photosensitive properties of Selenium to enable the blind to recognise simple objects in their surroundings. The elektroftalm consisted of a Selenium photocell placed on the forehead of a blind person, and a device that transforms light into sound. Using the device it would be possible for a blind person to detect the light from a window, lamp, and so forth. This device was later (1963) modified and converted into an 80-channel (points) vision-to-tactile device by W. Starkiewicz & T. Kuliszewski. Using the modified versions of the Elektroftalm, after two weeks of practise with the device, a blind user was able to identify very simple objects. [StaKul63, WarStr85 – pp.1-12]
Turine, V. de	1902	Photophonic Book	Reading	A reading machine that used specially prepared texts. Each letter on a page would be represented by a series of small transparent squares. By shining light at the page a series of flashes are detected and used to modulate an electric current. Thus producing a distinct sound for each letter of the alphabet. [CooGaiNye84]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Fournier d'Albe, E. E.	1912	Optophone	Reading/ Obstacle	He created the first optophone in 1912, the 'Exploring Optophone', which created various tones depending on the amount of light falling on a single Selenium cell. By 1914 he had created the 'White sounding' optophone which could be used to 'slowly' read print from a book. [Bar21, Bar30, Bed68, Beu47, CapPic00b, Cof63, CooGaiNye84, EleRev21, Fou14, Fou24, Nye70, SciAme20, WarStr85]
Brown, Prof. F. C.	1915	Phonopticon or Phonoptikon	Reading	Uses a system of three or four selenium crystals, instead of cells, due to the higher light-sensitivity. But relied upon the 'full effect' of the selenium crystals, instead of the 'flash effect' of the original optophone, so information loss can occur due to blending of the output signal. (Similar to the problem of human sight when moving from light to darkness – the eye requires a period of readjustment, as does the selenium crystal). [Bar30, Beu47, Fou24]
Messrs. Barr & Stroud, Ltd.	1918 - 1921	Optophone	Reading	In 1918, after the war, Barr & Stroud continued to improve the Optophone created by Fournier d'Albe, mainly by converting it to a 'black sounding' device. This was accomplished by adding a second selenium cell that is used as a balancer. One cell is continually exposed to the light, so that when white paper is in view of the second cell, the two opposing currents are equal in magnitude. Therefore no sound is heard, until a character is scanned and the current balance is broken. [Bar21, Bar30, Beu47, CapPic00b, CooGaiNye84, EleRev21, Fou24]
	1920- 1930	Guide Dog	Obstacle	First used in Austria as a guide for the blind. [Kay84]
Naumberg, Robert E.	1928	Visagraph	Reading	This device is a direct translating, non-integrating device employing a tactile output. The machine produced an enlarged, raised replica of the printed material by embossing aluminium foil. A page of text would be scanned by the machine in a similar fashion to the optophone (using Selenium detectors), which would directly control a series of embossing plungers that would move over the aluminium foil. However, to complete one page of text required in the region of 30 minutes, which was considered too long to be of use. (Improved in 1947 under sponsorship from the Committee on Sensory Devices). [Beu47, Cof63]
Haskins Laboratories' – Radio Corporation of America (RCA)	1944	RCA Recognition Machine	Reading	Scans a printed letter and compares its pattern with those stored in a matrix. When a match is found, a recording of that letter is played. Averaging about one letter per second. Later (1968>) Mauch Laboratories also created a 'Recognition Machine' [Lau68], which was the size of a typewriter, and could audibly spell out the scanned words, or display them using a Braille output. The machine could allow reading speeds of up to 90 words a minute. [CooGaiNye84]
Long, A. G	1944	Iconoscope	Reading	Prov. Patent Specification No. 2423 of 9 th February 1944 – 'Apparatus for enabling Blind Persons to Read and the Like'. A cathode ray tube with the normal fluorescent screen replaced by a mosaic of electrodes, each of which is connected to the corresponding electrode in a second mosaic, placed on the skin. An image is displayed on the first mosaic of electrodes, which then passes the signal onto the skin, via the second set of electrodes. (This was also suggested by Joseph Morgan Ward from South Africa, in Prov. Patent Specification No. 7816 of 17 th December 1946). [Beu47]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Radio Corporation of America (RCA)	1946	RCA Type A-2 machine	Reading	The Radio Corporation of America (RCA) created a direct translating, non-integrating device called the RCA Type A-2 machine which worked in a similar way to the original Optophone, but additionally employed a vertical sweeping spot of light and a variable frequency oscillator. The spot of light would scan a vertical section of a character, and if the light encountered a letter contour, a frequency would be generated. The frequency being determined by the vertical height of the beam when the contour was first encountered. [Cof63]
	1946>	Photoelectrograph	Reading	Similar to the Visagraph, in which a 42-element mosaic of plungers received each letter in turn from a scanned page of text. [Beu47]
Messrs. Tunncliffe & Wilkins	1947	Radioactive Guider	Obstacle	A suggestion was made for a guiding device that produced a 'beam of radiations' by using a 'small particle of radioactive material'. The radiation reflecting of nearby objects could be detected via a Geiger counter, and used as an indication of the object's distance. (How far they managed to get with these rather perilous trials can only be speculated upon!). [Beu47]
Cashin, J. A	1947>	Cashin Reading Machine	Reading	Unaware of the existence of the Optophone, this device was created as a 'black sounding' machine, with an output consisting of five musical notes. As with the 'black sounding' optophone, a type of balancer is used so that no sound is heard when viewing a blank white page. Using the device a speed of 15 words per minute could be obtained by Mr. Cashin himself. [Beu47]
Quentin, Prof. St.	1947>>	St. Quentin Viewing Machine	Reading	The device compares a number of masks of negatives of each letter of the alphabet with the characters on a page of text. When the correct mask is found, a minimum amount of light will pass through to a photocell, and so the relevant sound is played. (A Berlin engineer, Georg Schutkowski, patented a similar idea in 1922). [Beu47]
	1950-1960	Long Cane (Typhlocane) (Further development of the long cane)	Obstacle	During the 1950's a group of enthusiasts were directed towards the use of a non-electronic aid. By 1960 a university Master's level course was introduced at Boston College to train peripatologists (orientation and mobility specialists), who would then teach blind persons on a one-to-one basis to be independent travellers through the use of a 'long-cane travel technique'. The user continually moves the cane in an arc in their path, repeatedly tapping the ground or nearby objects, such as a fence. The user is then alerted to the presence of an object either by contact, or by the echo generated from the sound of the tapping cane. [ClaHeyHow86, Dod85, Kay84, Kay85]
Benham, T. A. Benjamin, J. M.	1951>	Optical Obstacle Detector	Obstacle	Pursuing the concept of the US Army Signal Corps for an optical obstacle detector, Benham and Benjamin created a device that used two narrow interacting pulsed ultraviolet beams and a sensor, which received the reflections from objects. [BenBen63]
Beurle, R. L.	1951>	Clicker	Obstacle	Produces audible 'clicks' that echo off any nearby obstacles. The subject can learn to interpret the echoes and determine their immediate surroundings. However, once a blind user became familiar with audible echoes resulting from the clicks, it became easier for them to use their own sounds, for example, to tap their feet, or a cane as they walked. [Kay84]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Kallman, H. E.	1954>	Optar	Obstacle	Used as a range finder. Sends out electromagnetic radiation in the infrared and visible spectrum. Detects range to an object by automatically calculating the distance behind a lens at which the image is in sharp focus. The pitch of the tone of sound produced was proportional to the distance to the nearest detected object. The tone of the sound output was generally considered to be too difficult to comprehend adequately, and the device couldn't detect curbs. [BenBen63, Jac63a]
Flanagan, Patrick	1958	Neurophone		While only 14 years old, Patrick Flanagan created the Neurophone (similar to a later device called the Grokbox), which converts sound waves into signals that approximately match those understood by the human brain. It is believed that these signals travel to the brain via the nerve pathways throughout the human body. In this way, the user (only) can hear sound through the device, such as music, by simply placing two electrodes from the device onto the skin. The use of this device for the blind is in its ability to present sounds to a listener without the need for headphones. Blind aids that utilise a sound output often prevent users from hearing sounds from the local environment (such as a nearby moving car). However, with the Neurophone a user could hear both the output from such a blind aid, as well as noise from the local environment. [Numerous references can be found on the internet for the Neurophone]
	1958>	Battelle (Optophone) Reader	Reading	Variation of the original optophone that allows for reading speeds between 10 and 15 words per minute. [Cof63, Fis76, Lip58]
Lipton, Arthur Prepared by Lipton for Veterans Administration.	1958>	Optophone	Reading	Variations of the original Optophone device, such as 'saw-tooth' and 'stair-step' devices. Devices using about eleven photocells appeared to be much more accurate than the five used in the original Optophone. The greater number of possible tones (i.e. 11 tones for 11 photocells) apparently aided in the training of new users. One problem noted with the device came from the varying conductivity of the photocells due to the heat emanating from the device itself. [Lip58]
Benham, T. A., Benjamin, J. M	1960>	Benham- Benjamin Infrared detector	Obstacle	Infrared obstacle and curb detector similar to the Optar. [BenBen63, Jac63a]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Benjamin, J. M., & Benham, T. A.	1962	Laser Cane (Nurion Cane)	Obstacle	A sensory aid, introduced by Benjamin, J. M. (amongst others), that received a great deal of support (financially speaking) from the Veterans Administration over a period of almost 30 years. A similar cane was also developed in Sweden a number of years later (about 1974). Pre-1960s ultraviolet beams were used, however this was changed to three lasers. These three lasers pointed down, ahead, and up and would provide information about any nearby obstacles. A rasping 200Hz tone indicated a possible step-down ahead. A vibrating stimulator against the index finger warned of objects 1.5m to 3.6m ahead. A high-pitched tone of 2600Hz indicated the presence of an object ahead of or well above the cane. By twisting the cane while walking information could be gathered about the user's surroundings. [BenAliSch73, BenBen63, Bra85, ClaHeyHow86, Fis76, Kay84, Spu85, WarStr85]
Johnson, Avery R.	1963	Stereo-Optical Edge Detector	Mobility	(Proposed device – completion unknown) Uses a stereo camera system, with a mosaic of photosensitive elements in place of the camera film, to locate difference boundaries. The device only locates objects that are 'relevant', near, to the user. But the device can only detect those edges that have a non-zero vertical component, unless the device is tilted. The device could be used to detect pathways, windows, doors, lights, and so on. [Joh63]
	1963>	Electrical Cane	Obstacle	Electrical canes were produced that emitted electromagnetic radiation, which enables a user to detect a curb without tapping the ground. But this doesn't supply a very effective early warning system, since the range is limited to a few inches farther than the length of the cane. [BenBen63]
	1963>	Gyro Compass		More accurate than the magnetic and radio compass, but its complexity makes it difficult for the blind to use. The device is intended to help guide a blind user to a known destination by allowing them to keep to a fixed direction. [Jac63b]
	1963>	Horn	Obstacle	Would be used by a newly blind person who had not yet achieved the required level of perception with their hearing to simply understand the echoes caused by a cane or footfall. The horn would then be used to produce a short, high frequency sound that would enable better judgement of nearby objects. Unfortunately, this method does not provide a sufficiently narrow beam for frequencies in the human audible range. [BenBen63]
	1963>	Radio Compass		Requires the use of a radio, tuned into a local radio station. The user holds the radio so that either a null, or a maximum is reached in the signal. Then by walking while keeping the radio signal at a constant it is possible to walk in an approximately straight line. (Later a device was created that produced a continual beat that grew in intensity as the radio station became clearer). [Jac63b]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Foster, David Blythe	1963>	Aural-Visual Devices	Mobility/Obstacle	A device was developed that could be incorporated into a handheld system or spectacles. The device captured images from a camera and produced sound in a similar manner to the optophone. One interesting feature was the use of coloured filters positioned in front of various photocells that allowed the user to identify colours from the scene by listening for different frequencies and musical notes. [Fos64]
Jacobson, Bertil	1963>	Ambient-Light Obstacle Detector	Obstacle	Similar to the Kallmann Optar device. The system uses a lens or concave mirror to focus a sharp image of an object. Once accomplished, the distance between the lens and the location of the 'sharp' image is a direct function to that of the distance from the lens and the object. The device uses a system of automatic ranging over a range of distances from 0.5 to 10.0 metres, and has a tactile display as an output. The device is inferior to the Benham-Benjamin infrared detector in the sense that it cannot identify objects, like walls, that have an even surface with little or no variations in contrast. But it does allow for a wider search area than the infrared device. [Jac63a]
Jacobson, Bertil	1963>	Magnetic Compass & Straight Course Indicator		The system allows a beam of light to shine onto a photoconductive cadmium sulphide cell when the compass needle is in the direction of the intended course. The resulting current is amplified and made to operate a relay-like mechanism causing a peg to protrude from the top of the box. Major distortions can occur when close to large metal objects, such as cars or steel girders. [Jac63b]
Starkiewicz, Witold & Kuliszewski, Tadeusz	1963>	Elektroftalm	Obstacle/Mobility	This later device, based on Noiszewski's 1897 original, converted light energy into tactile stimuli. The device uses a system of semiconductor photoelements that receives light, and produces an electrical pulse for each photoelement. The electrical pulses are then converted into mechanical energy via a mosaic of tactile elements placed on the skin, equal in number to that of the array of photoelements. The intensity of the tactile stimulation being a function of the intensity of the incident light. The forehead was used for the tactile display, originally having 80 photoelectronic channels, but later using 120. The maximum resolution of the device being in the order of 2 degrees, with a field of view of 28 degrees. [StaKul63]
Kay, L.	1965	Sonic torch	Obstacle	The first ultrasonic blind aid developed by Kay. Hand-held device with an audible display presented through a single hearing aid-type earphone. The output is a pulsed tonal complex in which pitch rises linearly with distance. The sonic torch transmitted a wide-bandwidth (40-80 kHz), frequency-modulated, ultrasonic energy wave. Reflected signals would be converted to the audible region by multiplication with the transmitted signal. The pitch and timbre corresponding to the range and variation in target surface texture, respectively. The latter enabling some degree of target identification. This device was the first spatial sensor for the blind to become commercially available, and by 1968 there were approximately 1000 of these devices in use. [Bra85, Kay84, WarStr85 – pp.1-12]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Kay, L.	1965	Sonicguide (Binaural sensory aid)	Obstacle	<p>Kay's original sonic devices were introduced in early 1965 and used microprocessors and ultrasonic units. The system works by using a downswep FM ultrasound signal that is emitted from a transmitting array with broad directional characteristics in order to detect obstacles. A two-channel receiver picks up the reflected signals. The frequency of the ultrasound being swept from 70 to 40kHz within 1ms. So having similar characteristics to that of a bats echolocation system. The detected signals are then converted to a binaural audio signal. The device has a very good resolution, being able to detect objects such as a wire 1mm in diameter, or discriminate between several adjacent objects. The FM-downsweep was found to be superior to the upsweep method for locating small objects, and that the sweeping ultrasound signal allowed for a greater accuracy than a constant frequency signal when determining distances and object sizes.</p> <p>It was found that the device could not easily be used with the long cane. So the device was designed into the originally planned spectacle frames (see - Ultrasonic Spectacles). The new device, which used a wider scan range in the order of 55-60°, could be used to detect complex objects in the user's surroundings. It was also demonstrated that with training a user could focus their auditory attention to one individual object out of many in the local environment.</p> <p>See also the 'Ultrasonic Spectacles' by L. Kay (1965). [Bra85, Kay84, Kay85, ShoBorKor98, WarStr85 – pp.1-12]</p>
Kay, L.	1965	Ultrasonic Spectacles (See Sonicguide)	Obstacle	<p>Derived as an improvement to L. Kay's original Sonicguide (1965). Radiates ultrasonic energy, which is reflected off obstacles. A binaural system is used, with the left-right directions coded primarily by the intensities of the signals in the two ears. The output frequency is varied according to the relative distance to a detected object, and the amplitude and pattern of the sound is varied by the objects reflective properties. The sound is presented to the ears via small plastic tubes, which do not touch any part of the ears. In this way, sounds appear to originate from a location in or around the head, which in conjunction with the intensity of the sound heard, can give the approximate location of any obstacles. [BenAliSch73, Bra85, Fis76, Kay84]</p>
Russell, Lindsay	1965◊	Russell Pathsounder	Obstacle	<p>One of the earliest ultrasonic travel aids. The device uses a 30-degree ultrasonic beam transmitted from two ultrasonic transducers on a chest-level unit suspended from the user's neck. The unit gives three levels of feedback in the form of tactile or auditory (clicks) warning of objects up to six feet ahead allowing for limited range estimation. [BenAliSch73, Bra85, WarStr85]</p>

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Bliss, J. C., Linvill, J. G. Manufactured by Telesensory Systems, Inc.	1966>	Optacon	Reading	<p>The Optacon reached the market in the early 1970's, and could read printed material, or even text from a computer screen. The device is a form of TVSS (Tactile Vision Substitution System).</p> <p>The device produces the shape of printed letters, appearing under a camera, on an array of 6x24 vibrating pins. The camera is manually moved over the text, while the other hand is placed with the index finger in a groove over the vibrating pins allowing the user to read the text.</p> <p>Training time is normally in the region of two weeks. The device can be used to read almost any printed material. Reading speeds of 80 words per minute can be obtained, but 40 is average. The Optacon is fairly difficult to learn to use, with most people setting it on a par with learning a new language.</p> <p>Note – the output of the Optacon is not Braille, rather it is the approximate tactile image of what is scanned. The device 'sees' contrasts on the page being scanned, and produces a vibrating pattern to match. This implies that the device is a 'direct-translation' reading aid.</p> <p>[BacHug85, Fis76, Kay84, Lau89, Moo86]</p>
Collins, C. C., & Bach-Y-Rita, P.	1968	Tactile television	Obstacle/ Mobility	<p>A matrix of 256 electrodes built into a body belt was constructed together with a waistcoat containing the electronics and battery power supply coupled to a solid-state camera mounted in spectacle frames for wearing on the head. The idea of electrocutaneous stimulation didn't seem to hold the answer to the blind mobility problem since the useful field of view was either too limited or the discrimination of objects against the distant background over a wide arc was inadequate. Other problems being caused by the occasional lack of electrical contact, which could occasionally lead to slight burning, giving rise to the need of a <i>dead man's switch</i> allowing for a rapid shut-off of the system.</p> <p>It was originally believed that electrocutaneous stimulation could provide a grey scale that couldn't be achieved with normal tactile systems, but the idea didn't work well in practise due to the above problems.</p> <p>[BacColScaHolHar70, Col85]</p>
Mauch Laboratories	1968	Visotactor	Reading	<p>Uses eight photocells generating signals to activate eight tactile stimulators. Two stimulators on each finger of one hand. Allowing a reading speed in the order of 15 words per minute. [Fis76, Lau68]</p>

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Dobelle, W. H., Quest, D. O., Antunes, J. L., Roberts, T. S., Girvin, J. P. G. (Many others have investigated this area from the late 1960s onwards)	1968<	Electrical stimulation of the visual cortex	Mobility	Originally pioneered by Giles Brindley, and later by William Dobelle et al. An image on the camera retina is transferred to an array of electrodes implanted on the surface of the visual cortex. Each electrode, when energised, generates a phosphene. When several electrodes are generated, a pattern is perceived. But before this pattern has spatial meaning, however, the cortex has to be mapped so that a direct translation from 'retinal' image to perceived pattern takes place. The prospects seem limited though, since the mapping is a tedious and difficult process, which appears to be a serious obstacle to overcome. The advantages currently don't justify the trauma of the implants, and the apparently difficult learning process (although this may change with continuing advancements in both technology and medical techniques). [Kay84, Mei00] Similarly, around 1977 research was undertaken investigating the implantation of electrodes inside the cranium of several sightless volunteers, to stimulate the visual cortex. The subjects were able to perceive simple black and white pictures. [Smi77]
	1968>	Talking books	Reading	Cassette recordings of a sighted reader reading aloud text. [Bed68]
Beddoes, Michael, P.	1968>	Optophone	Reading	A number of improvements were made in the 'new' optophone. For example, using time-compressed signals to improve the comprehension and learning of sounds generated by the device that correspond to letters and words from the page of text being scanned. It was also believed that the musical tones were generally less strenuous than multidimensional and multimodal coding schemes. [Bed68]
Mauch Laboratories	1968>	Visotoner	Reading	Allows for reading speeds between 10 and 15 words per minute. Later redesigned, encoding signals in a similar manner to the Battelle Optophone (Battelle Reader), with an average reading speed of 40 words per minute. Uses a column of nine photocells, producing a different audible tone for each photocell while it is 'seeing' black. [Fis76, Lau68, Nye70]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Bach-Y-Rita, P. Collins, C. C.	1969<>	TVSS (Tactile Vision Substitution System)	Obstacle/ Mobility	Tactile Vision Substitution uses a two-dimensional 400-point matrix of stimulators (commonly placed over the abdomen) to present spatial information. Unfortunately, due to the low resolution obtainable, the devices generally cannot be used as a mobility aid. But the devices can aid recognition of simple objects. Although the original TVSS could be used to detect and recognise simple objects, like a telephone or toy horse, the objects had to be against a plain high-contrast background. The other limiting factors were the 5-degree scanning field of the camera, and the overall size of the device, which was generally considered to be too bulky. But Bach-Y-Rita & Collins, along with Linvill and Bliss with their reading machine, which lead to a marketable device called the Optacon, became some of the main pioneers of the vibrotactile display. The Optacon itself was later modified to combine both auditory and tactile outputs to improve its usability. [BacColScaHolHar70, BacHug85, FriBacTomWeb87, KacWebBacTom91, Mei00, WarStr85-pp.1-12]
	1970>	Cognodictor	Reading	Uses a method of letter recognition in a similar way to that of the Lexiphone, generating a different sound for each letter scanned. [Fis76]
	1970- 1991	Sonospec	Obstacle	The Sonospec is a 'clear path indicator' or 'obstacle detector' and alerts the user to the presence of a nearby obstacle, such as the Sonic-torch, Pathsounder, Mowat Sensor, Nottingham Obstacle Detector, or Laser-Cane. It works in much the same way as a torch is used in the dark, with a narrow beam that provides information on one point only. [IfuSasPen91]
Beddoes, M. P., (Suen, C. Y.)	1971>	Lexiphone	Reading	Printed text is scanned using a column of 54 photodiodes. Each letter of the alphabet having its own distinctive sound. Dichotic presentation was incorporated in the device output, resulting in much quicker reading speeds with the general rate being in excess of 30 words per minute. [BedSue71, Fis76]
Fish, Raymond M.	1972	Audio display	Mobility/ Obstacle	By 1976 four types of machine had been created, two of which used an array of photocells, the others utilising a video camera, to produce a series of pulses of sound. The systems worked only as edge detectors, so in principle, two levels of sound. All edges producing a sound, elsewhere nothing is heard. The method used for coding the images employed the principles of psychoacoustics in that high pitched tones naturally seem to come from a high location, and amplitude differences in the sound presented to the ears make it seem that sounds are coming from certain positions from left to right. Test subjects could rapidly detect and recognise objects and shapes. One drawback of the device came from the difficulty in accurately determining distances to objects that are over five feet away from the device. [Fis73, Fis76, Ohe94]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
	1972>	Stereotoner	Reading/ Mobility	Variation of the original optophone, using a system of ten photocells to convert light into electric current. This is then converted into musical notes with (originally) a frequency range of 440-3520Hz. Lower tones are generated by the lower parts of letters and heard through the left earphone; higher tones are generated by the top half of the letter and heard in the right ear. The device can also be used as a simple mobility aid. The Stereotoner can be used to reach speeds equal to that of the Optacon (40+ words per minute), but people tend to find this device harder to learn. Although, in practise the Stereotoner is more versatile and can be used to a slightly greater extent than the Optacon. [Lau89, VerWan85]
Lauer, Harvey	1973- 1989 (Prob. Mid- 1980s	Optaudicon	Reading	After evaluating the Optacon and the Stereotoner, it was found that the Stereotoner had a superior optical system. But the Stereotoner's photosensitive elements and vertical resolution were very much inferior. The Stereotoner could greatly be improved by increasing the 10-tone code to a 20-tone code. By combining this new tonal output with the Optacon, and re-christening the device, the 'Optaudicon', it was found that the combination of auditory and tactile outputs could be used in conjunction to aid recognition of printed material. While some people would prefer to use only one particular output, it was demonstrated that the bimodal approach would have a complimentary effect. For example, tactile code has a wider window for tracking; it is also better for vertical resolution; but the finger loses sensitivity (feels numb) more quickly than the ear. The tonal code is better for reading dense horizontal data and descenders, etc. Hearing is not as susceptible to such loss of sensitivity or negative adaptation as is touch. [Lau89]
Kurzweil Computer Products Inc.	1975	Kurzweil Reading Machine (KRM)	Reading	Several versions of the KRM (at the time costing - \$20000 to \$30000) were created, followed by a more advanced device called the Kurzweil Personal Reader (cost - approx. \$10000). The Kurzweil reading machine, which is or was found in some libraries, could be wheeled round to a table, then the book to be read would be placed face down on the camera unit. The machine could then read (speak) the book, with various keys to control the operation of the machine. The machine had three stages in converting print to speech. Recognition and analysis of the printed material. Conversion of the character strings to phonics. Generation of the speech. If a word was not recognised, and so could not be spoken, the machine would pronounce the individual letters. [Lau89, Kay84, Moo86]
Wormald Sensory Aids International Ltd.	1977>	Mowat Sensor	Obstacle	Hand-held ultrasonic torch, which is normally carried in a pocket or purse, and used to locate objects that are out of reach of the long cane. It radiates short ultrasonic pulses and operates on the 'sing around' method of determining distances. An echo from the nearest object within a maximum range of 4m and which exceeds a set threshold re-triggers the transmitter. Vibration of the torch indicates the presence of an object in its elliptical beam of 15 degrees azimuth and 30 degrees elevation. A slow vibration of 10 Hz for an object at 4m increases inversely proportional to distance up to 40 Hz for an object at 1m. The device is fairly cheap, and very easy to use. This device was designed to be a secondary aid to be used in conjunction with the long cane or a guide dog. [Bra85, Kay84]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Kay, L. Strelow, E. R.	1979>	Canterbury babies' aid	Obstacle	A modified version of L. Kay's Binaural Sonicguide. After testing the device on a blind baby from the age of 9 months, through to the age of three and a half years, it was found that the child had learnt to use the device in such a way that they could behave more like a normal sighted child in the sense that they had gained spatial awareness that is often lacking in people who have been blind from birth. [HilDodHilFox95, Kay84]
Kay, L. Strelow, E. R.	1979>	Canterbury child's aid	Obstacle	Similar to the Canterbury babies' aid, but with different settings for the range, and resolution, and so on. The device was a form of binaural sonic guide. [HilDodHilFox95, Kay84]
Dallas, Stanley A., Jr. Thales Resources Inc., US	1980	Sound Pattern Generator	Mobility/ Obstacle/ Reading	In Dallas' device, vertical position in the scene was mapped to frequency, horizontal position to time, and brightness to loudness. The mapping segmented a 2-D image into vertical strips that were processed and displayed audibly, scanning from left to right. This form of mapping is an example of the piano transform. [DalEri80, Ohe94]]
	1980<>	Nottingham Obstacle Detector	Obstacle	Hand-held ultrasonic torch, similar to the Mowat sensor, which radiates short pulses of ultrasonic energy in a narrow beam. The nearest object in the beam to produce an echo, which exceeds a threshold, is indicated by one of eight possible tones from a miniature loudspeaker, the tones corresponding to notes of the major musical scale. The maximum range is 210cm, with the musical note increasing in 30cm steps. (The device was still not commercially available in 1984). [Bra85, Hey85, Kay84]
Brabyn, J. A., Collins, C. C., Kay, L.	1981>	Wide bandwidth scanning sonar	Obstacle/ Mobility	Wide bandwidth scanning sonar with tactile and acoustic display for persons with impaired vision (the blind, divers in dark murky water, etc.). The device used a wide-bandwidth variable-speed electronically scanned air sonar, based upon a system by Kay, coupled to a vibrotactile display. The spatial information, presented as a plan position indication of objects up to 5m, allowed for better mobility. Unfortunately at the time the scanning sonar and tactile display was too cumbersome and technologically over-sophisticated for commercial application to the blind. [Col85, Kay84]
Kurcz, E.	1981>	Heliotrope	Mobility	Kurcz developed a hand-held device that utilised a point mapping, which sensed the light output from only one point in the scene. The sound output of the device is a function of the light intensity captured. [Ohe94]
	1982>	Polaroid ultrasonic travel aid	Obstacle	Similar to the Mowat sensor and Nottingham obstacle detector.

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Kay, L	1983>	Trinaural sensory aid or Trisensor (KASPA)	Obstacle	The Trisensor, also known as the KASPA system (Kay's Advanced Spatial Perception Aid), is based on the Canterbury devices. Has the same ability as the binaural devices, but adds a narrow high-resolution central beam over the left and right binaural channels to gain information to greatly improve object identification. The device demonstrated a threefold improvement in resolution over the binaural device. It became apparent that if the device were given to a blind baby or infant, allowing them to use the device over a period of some months or even years, they could learn to use the device to a greater extent than most adults could. In fact, even to the point that they could use the device and accomplish some tasks (in terms of spatial location and awareness) that were comparable to a person with partial sight. [BriGra87, Kay84, Mei00]
Heyes, Tony	1984	Sonic Pathfinder	Obstacle	A modification of the Nottingham Obstacle Detector. Similar to Kay's Trinaural sensory aid, it uses a head-mounted pulse-echo sonar system comprising three receivers and two transmitters. Unlike Kay's device, the auditory output tends to only provide information on the nearest object, and it does not provide information about surface texture. [ClaHeyHow86, Dod85, Hey85, Mei00, WarStr85-pp.29-34]
	1984>	Paperless Brailler	Reading/ Writing	A form of word-processor for the blind. Storage medium, at the time, being C60 cassette tapes. The device uses a row of electromechanical Braille character cells acting as a line of text. The line of Braille text can be scanned with the fingers, allowing both writing and reading of whole pages of text. [Kay84]
	1984>	Perkins Brailler	Writing	Typewriter for Braille. [Kay84]
Televoid Systems Ltd.	1984>	Closed-circuit TV systems	Reading	TV camera is used to magnify and display text or pictures on a television screen. The system is used as a low vision aid for people whose visual acuity is deteriorating. [Kay84]
Triformation Systems Inc. (Amongst others)	1984>	Braille Printer	Writing	The machine prints sheets of Braille writing from punched paper tape, containing Braille code, obtained from a special typewriter or from a digital signal. [Kay84]
Wormald Sensory Aids International Ltd.	1984>	Viewscan	Reading	A line of 45 optical fibres mounted in a hand-held camera is scanned across the print in a continuous motion. The motion of the camera over the page is transferred electrically to the machine, like a 'mouse' with a computer. Two microcomputers would then process the image signals received from the fibre-optic cable, and produce the picture on a flat, thin, 45 x 115 matrix, low-resolution plasma display. The device can display the text with a number of different magnifications, and is specifically aimed towards those with poor visual acuity. Unfortunately it is generally considered that the orange appearance of the letters makes for difficult reading for many users. [Kay84]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Nielson, Lars. Mahowald, Misha. Mead, Carver	1987>	SeeHear	Mobility	The SeeHear device calculates the position of light sources in 2-D, and the light intensity information. Synthesising a sound having the appropriate psycho-physiologically determined cues for a sound source at the specified location. Since the device emphasises changes in light, if a bird were to fly past, a sound would be generated that would correspond to the location, brightness and motion of that bird. The device has been created to encode the audio signals with 3-D information, so that the bird in the above example could be located in space due to its movement. [NieMahMea87]
	1988	Optacon II	Reading	The Optacon device repackaged in a modular fashion, with a computer interface. The device has been built into a larger reading machine providing slightly superior performance over the original Optacon. [Lau89]
Borenstein, J., et al.	1994	NavBelt	Mobility/ Obstacle	The NavBelt is a portable device equipped with ultrasonic sensors worn on a belt around the waist, and was originally conceived of in 1989, and first built and tested in 1994. The belt contains eight sensors covering a combined field of view of 120°. The system provides feedback about the location of either obstacles in the users path or the direction of a clear path in the form of binaural auditory cues, based on interaural time difference (phase difference between left and right ears) and amplitude difference. [BorUlr97, ShoBorKor98]
Meijer, Peter	1996	The vOICe	Mobility/ Obstacle/ Reading	In 1991 'the vOICe' was created in the form of a hardware device, and then later in 1996 a software version of this video optophone was developed. Images are captured via a video camera and relayed to a computer. The captured image is then reduced to 64x64 pixels and converted to sound using stereo effects (left part of image is heard in left ear, and so on). The resulting sound is also encoded so that it uses 16 different amplitudes for different pixel intensities, and uses low frequency sounds to represent the bottom of the image, and high frequency sounds for the top. (Note - 'the vOICe' can be downloaded from Peter Meijer's website [Mei00]) [Mei00]
ThaleScope Ltd.	1997	ThaleScope	Mobility/ Obstacle/ Reading	Testing has begun on a device called the ThalesScope, which is supposed to convert sight into sound. The device appears to be almost identical to Peter Meijer's 'the vOICe', however, as of yet little information is known. [Mei00]
Borenstein, J., et al.	1997>	GuideCane	Obstacle	The GuideCane can almost be thought of as a robotic guide dog. It is held out in front of the traveller in a similar fashion to that of the white (long) cane, but it has a series of ultrasonic sensors housed in a unit at the end of the cane, which is supported by two guide wheels. This unit also contains a steering servomotor that is operated by an onboard computer. When an obstacle is detected the device calculates a new course that manoeuvres the user around the obstruction. [BorUlr97]

Name	Year	Device(s)	Use	Description of the device, and its proven/possible capabilities
Querelle, Paul Anglia Polytechnic University	1998	Camsight	Mobility/ Obstacle/ Reading	Work began in 1998 at Anglia Polytechnic University in Cambridge England with a device known as Camsight. Areas of interest in the research involve: aspects of vision to sound, extracting shape information from images, extracting depth information from multiple images, locating a mouse pointer using sound to facilitate menu navigation, use of the inverse Fourier transform with windowing to construct complex soundscapes, determination of texture in an image to facilitate OCR and the use of reverberation and other filters to enhance auditory aesthetics. [Mei00]
Reid, Harry	1998	Sonic Eye	Mobility/ Obstacle/ Reading	The system appears to work in a similar manner to Peter Meijer's 'the vOICe' converting images into sound. [Mei00]
Rohrhuber, Julian & Wittchow, Oliver University of Hamburg	1999	NanoVoice		A combination of Peter Meijer's 'the vOICe' with a Nintendo GameBoy. [Mei00]

Appendix 2. Test data

A2.1. Test 3 – Preliminary test data

The results that follow were obtained from the first 7 subjects during the preliminary tests comparing the cartoon depth map with plain unmodified images looking for accuracy in the subjects' ability to perceive depth and to determine the location of two large letters ('A' & 'B') within a scene.

All seven subjects had a ten minute training period followed by the test, consisting of 16 images – 8 plain unmodified, and 8 showing the same scenes in the form of cartoon depth maps. The tests were supervised and all comments noted.

The data file for each subject has the following appearance:

	Y/N	Type	LD	LD		Repeats required		
1	aB	N	NY	NN	ba	5		
2	bA	N	NN	NY	AB	7		
3	bA	N	NY	NN	ab	6		
4	ba	Y	C	YY	YY	3		
5	AB	N	YN	YY	aB	3		
6	aB	Y	C	YY	YY	6	3→Far Near	
7	Ab	Y		YY	YY	6	5→Letters correct	
8	AB	N	C	NY	NY	BA	5	3→Near Near
9	ab	N		NN	NY	Ba	4	1→Decided far far
10	ab	Y	C	YY	YY		4	2→Far Far
11	Ab	N	C	NY	NY	Ba	4	2→Near Far.
12	bA	N		YN	YY	BA	4	1→Letters correct
13	bA	Y	C	YY	YY		2	1→Far Near
14	Ab	Y	C	YY	YY		4	
15	bA	Y		YY	YY		4	2→Letters correct
16	AB	Y	C	YY	YY		8	3→Near Near

The first column displayed represents the frame number. The second is the subject's choice for that particular frame, with a lowercase character indicating a distant letter and an uppercase character representing a near letter within the scene. Similarly, the left and right positions of the characters in the scene are denoted by the order of the letters shown in the second column (i.e., aB → 'A' left and far, 'B' right and near). The image type represents what kind of image was used ('C' for cartoon, otherwise a plain image). 'LD LD' is used to represent whether the subject correctly determined the left Letter (L) and Depth (D) and similarly for the right Letter (L) and Depth (D).

If the subject's choice was incorrect then the correct letter positions (labelled the 'Actual image') are listed in the notation described above (i.e., aB). Finally, the number of repetitions required for the subject to determine the scenes are listed along with any comments that the subject made or any observations made by a supervisor. For instance, in the example above the comment for frame 6 is '3→Far Near', which implies that the subject believed after three repetitions that the left letter was far and the right character near. The subject was unable to decide on the order of the characters until the sixth repetition, whereupon they chose 'aB', which happened to be correct.

Subject 1

After 10 minutes practice listening to the 2 second scans/frames.

L=Letter, D=Depth, C=Cartoon, Depth Map, P=Plain image, Y=Yes, N=No

	Y/N	Type	LD	LD		Repeats required
1	aB	N		NY NN	ba	5
2	bA	N		NN NY	AB	7
3	bA	N		NY NN	ab	6
4	ba	Y	C	YY YY		3
5	AB	N		YN YY	aB	3
6	aB	Y	C	YY YY		6 3→Far Near
7	Ab	Y		YY YY		6 5→Letters correct
8	AB	N	C	NY NY	BA	5 3→Near Near
9	ab	N		NN NY	Ba	4 1→Decided far far
10	ab	Y	C	YY YY		4 2→Far Far
11	Ab	N	C	NY NY	Ba	4 2→Near Far.
12	bA	N		YN YY	BA	4 1→Letters correct
13	bA	Y	C	YY YY		2 1→Far Near
14	Ab	Y	C	YY YY		4
15	bA	Y		YY YY		4 2→Letters correct
16	AB	Y	C	YY YY		8 3→Near Near

Cartoon 6 out of 8 correct

Depth 16 out of 16

Character 12 out of 16

$3+6+5+4+4+2+4+8=36$ repeats to recognise 8 frames.

Plain 2 out of 8 correct

Depth 10 out of 16

Character 8 out of 16

$5+7+6+3+6+4+4+4=39$ repeats to recognise 8 frames.

Subject 2

After 10 minutes practice listening to the 2 second scans/frames.

Subject is a musician.

L=Letter, D=Depth, C=Cartoon, Depth Map, P=Plain image, Y=Yes, N=No

	Y/N	Type	LD	LD		Repeats required
1	Ba	N C	YY	YN	BA	3 (1→letters correct 2→Cartoon)
2	Ab	N	NY	NY	Ba	6 (3→AB +3→Changed to Ab)
3	ab	N	YN	YN	AB	2 (2→ab parallel)
4	BA	N C	NY	NY	AB	4 (2→parallel)
5	ba	Y C	YY	YY		3 (2→far far)
6	BA	Y	YY	YY		3 (1→near near)
7	AB	N	YN	YN	ab	2 (1→parallel, thought near near)
8	ab	N	NY	NY	ba	4 (2→far far)
9	Ba	Y C	YY	YY		2 (2→Ba)
10	ab	Y C	YY	YY		2 (1→far far)
11	Ab	Y C	YY	YY		2
12	aB	Y C	YY	YY		3
13	aB	N C	NY	NY	bA	2
14	Ba	N	YN	YN	ba	2 (1→Thought Near for left)
15	AB	N	YY	YN	Ab	2
16	Ab	N	YN	YN	aB	2

Cartoon 5 out of 8 correct

Depth 15 out of 16

Character 12 out of 16

$3+4+3+2+2+2+3+2=21$ repeats to recognise 8 frames.

Plain 1 out of 8 correct

Depth 7 out of 16

Character 12 out of 16

$6+2+3+2+4+2+2+2=23$ repeats to recognise 8 frames.

Subject 3

After 10 minutes practice listening to the 2 second scans/frames.

L=Letter, D=Depth, C=Cartoon, Depth Map, P=Plain image, Y=Yes, N=No

	Y/N	Type	LD	LD		Repeats required	
1	bA	N C	NY	NY	aB	2	
2	AB	N	NY	NY	BA	2	
3	BA	N	NN	NY	aB	4	4→Thought was cartoon.
4	Ab	N C	NY	NY	Ba	2	
5	AB	N	YN	YN	ab	3	
6	AB	N	YY	YN	Ab	2	→Said sounds same as previous.
7	AB	Y	YY	YY		4	→Said sounds same as previous.
8	ab	Y C	YY	YY		2	
9	Ab	Y C	YY	YY		2	
10	aB	N C	NY	NY	bA	1	
11	ab	N C	NN	NN	BA	4	
12	AB	Y C	YY	YY		2	
13	AB	N	NN	NN	ba	2	
14	Ab	N	NY	NY	Ba	2	
15	BA	N	YN	YY	bA	2	
16	ab	N C	NY	NY	ba	2	→Thought was Plain image

Cartoon 3 out of 8 correct

Depth 14 out of 16

Character 6 out of 16

$2+2+2+2+1+4+2+2=17$ repeats to recognise 8 frames.

Plain 1 out of 8 correct

Depth 9 out of 16

Character 8 out of 16

$2+4+3+2+4+2+2+2=21$ repeats to recognise 8 frames.

Subject 4

After 10 minutes practice listening to the 2 second scas/frames.

Musician.

L=Letter, D=Depth, C=Cartoon, Depth Map, P=Plain image, Y=Yes, N=No

	Y/N	Type	LD	LD		Repeats required
1	ba	Y C	YY	YY		5 (2→far far 4→ab 5→ba)
2	Ab	Y C	YY	YY		2 (1→ab→C)
3	AB	Y C	YY	YY		2
4	Ab	Y	YY	YY		3 (2→ab→Near ???)
5	ba	N	YY	YN	bA	4 (1→ba)
6	ab	Y	YY	YY		3 (2→ab)
7	bA	Y C	YY	YY		3 (2→ba)
8	AB	N	YN	YY	aB	4 (2→ab Thought P, C, P, then finally C)
9	BA	Y C	YY	YY		5 (1→C 3→ba)
10	ba	Y	YY	YY		6 (5→ba)
11	bA	N C	NY	NY	aB	1 (Started to struggle from this point on)
12	ab	N	YN	YN	AB	3 (1→ab 3→P)
13	ba	N	YN	YY	Ba	2 (2→ba)
14	ab	Y C	YN	YN		2 (1→far far)
15	BA	Y	YY	YY		2 (1→ba→P)
16	Ba	Y C	YN	YN		3 (1→near far→C)

Cartoon 7 out of 8 correct

Depth 16 out of 16

Character 14 out of 16

$5+2+2+3+5+1+2+3=23$ repeats to recognise 8 frames.

Plain 4 out of 8 correct

Depth 11 out of 16

Character 16 out of 16

$3+4+3+4+6+3+2+2=27$ repeats to recognise 8 frames.

Subject 5

After 10 minutes practice listening to the 2 second scans/frames.

L=Letter, D=Depth, C=Cartoon, Depth Map, P=Plain image, Y=Yes, N=No

	Y/N	Type	LD	LD		Repeats required
1	bA	N C	NY	NY	aB	2 (1→far near)
2	ab	Y C	YY	YY		3 (2→ab→C)
3	AB	Y	YY	YY		1
4	BA	Y	YY	YY		2 (1→ab)
5	ba	N C	NN	NY	Ab	3 (1→ba→C→said 'b is far')
6	bA	N C	YN	YY	BA	2
7	AB	N	NN	NY	bA	2 (1→near near)
8	BA	N C	YN	YN	Ba	1
9	aB	N C	NY	NY	bA	1
10	ba	Y C	YY	YY		2 (1→far far)
(Said starting to get confused at this point)						
11	Ba	N C	NY	NN	AB	1
12	BA	N	NY	NN	Ab	2 (1→near near→P)
13	BA	N	NN	NN	ab	2
14	BA	N	NN	NY	aB	2 (Was unsure on type - decided on C)
15	BA	N	YN	YN	ba	3
16	Ab	N	NY	NY	Ba	3 (1→Ab 3→P)

Comments: Said that the cartoon images where easier to listen to.

Cartoon 2 out of 8 correct

Depth 12 out of 16

Character 8 out of 16

$2+3+3+2+1+1+2+1=15$ repeats to recognise 8 frames.

Plain 2 out of 8 correct

Depth 9 out of 16

Character 6 out of 16

$1+2+2+2+2+2+3+3=17$ repeats to recognise 8 frames.

Subject 6

After 10 minutes practice listening to the 2 second scans/frames.

L=Letter, D=Depth, C=Cartoon, Depth Map, P=Plain image, Y=Yes, N=No

	Y/N	Type	LD	LD		Repeats required
1	BA	N	YY	YN	Ba	3 Thought was Cartoon.
2	ab	N	YN	YN	AB	4 Thought was Cartoon.
3	Ab	N	NY	NN	BA	6 5→ab.
4	bA	N C	YN	YY	BA	2 1→C
5	Ba	Y C	YY	YY		3
6	Ba	N C	NY	NN	AB	4 4→Thought was plain image
7	ab	Y	YY	YY		3
8	bA	N C	NY	NY	aB	4
9	Ab	Y C	YY	YY		3
10	ab	Y C	YY	YY		3
11	AB	N	NN	NY	bA	3
12	ab	N C	NY	NY	ba	3
13	Ab	Y	YY	YY		2
14	ab	N	NY	NY	ba	3
15	AB	N	YN	YY	aB	3
16	bA	Y C	YY	YY		2

Cartoon 4 out of 8 correct

Depth 14 out of 16

Character 10 out of 16

$2+3+4+4+3+3+3+2=24$ repeats to recognise 8 frames.

Plain 2 out of 8 correct

Depth 10 out of 16

Character 10 out of 16

$3+4+6+3+3+2+3+3=27$ repeats to recognise 8 frames.

Subject 7

After 10 minutes practice listening to the 2 second scans/frames.

Musician.

L=Letter, D=Depth, C=Cartoon, Depth Map, P=Plain image, Y=Yes, N=No

	Y/N	Type	LD	LD		Repeats required
1	aB	N	YN	YY	AB	2 1→ab
2	ab	N C	NY	NY	ba	1
3	Ab	Y C	YY	YY		2 1→C
4	bA	Y C	YY	YY		2 1→C
5	aB	Y	YY	YY		3 1→P far far, 2→far close
6	AB	N	NY	NY	BA	2 1→P
7	bA	N C	NY	NY	aB	2 1→C far close
8	AB	Y C	YY	YY		3 1→C, 2→(ab)
9	ab	Y	YY	YY		3 1→P, far far
10	ab	N	NY	NY	ba	3 1→P, 2→(ab)
11	ab	Y C	YY	YY		1
12	Ba	Y C	YY	YY		2 1→C near far
13	bA	Y	YY	YY		3 1→P, 2→(ba)
14	Ab	N	NY	NY	Ba	3 1→P, 2→(ab)
15	ab	N	YN	YY	Ab	1
16	ba	N C	YN	YN	BA	3 1→C far near

Cartoon 5 out of 8 correct

Depth 14 out of 16

Character 12 out of 16

$1+2+2+2+3+1+2+3=16$ repeats to recognise 8 frames.

Plain 3 out of 8 correct

Depth 14 out of 16

Character 10 out of 16

$2+3+2+3+3+3+3+1=20$ repeats to recognise 8 frames.

A2.2. Test 4 – Test data

The results that follow were obtained from 20 new subjects during the second stage of testing, comparing the cartoon depth map with plain unmodified images looking for accuracy in the subjects' ability to perceive depth and to determine the location of two large letters ('A' & 'B') within a scene.

Each test consisted of two parts. First the subject would be trained (during a two minute training period) and tested with one particular type of image before proceeding, after a break, to the next type of image. Half of the subjects were tested with the cartoon depth images first, whilst the other half were trained and tested with the plain images first.

The same images as before were used, consisting of the letters 'A' & 'B' in the eight possible locations corresponding to *near* and *far*, and the different combinations of *left* and *right*. The test programme was fully automated, displaying the optophonic sounds (corresponding to the eight images) in a random order, with two randomly chosen repeats to provide a total of ten frames. On completion of the test, the programme wrote the data to an output file.

In the following lists of recorded data each frame is listed by number and then by the name of the file (indicating the characters location within the scene by the order of the letters and their respective distance from the camera in centimetres – 75cm or 150cm). Next the subject's selection is presented in the same form as previously described (i.e., 'Ab' – 'A' near left & 'B' far right), as well as the number or repetitions required by the subject.

Finally, the results are presented in a tabular form to provide a quick and easy way of perceiving the subject's success rate.

Subject 1 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 6 min.

1	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 3
2	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 5
3	A75cm_l - B150_r - cartoon depth	B a	Repetitions = 3
4	A75cm_l - B75_r - cartoon depth	B a	Repetitions = 5
5	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2
6	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 2
7	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 1
8	A150cm_l - B75_r - cartoon depth	a B	Repetitions = 2
9	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 2
10	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 1

Total number of repetitions is 24.

	CD	CD	Image	Reps
1	NY	NY		1
2	YY	YY	Y	5
3	NY	NY		3
4	NY	NN		5
5	YY	YY	Y	2
6	YY	YY	Y	2
7	NY	NY		1
8	YY	YY	Y	2
9	YY	YY	Y	2
10	YY	YY	Y	1
<hr/>				
Tot	6 10	6 9	6	24

Subject 1 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 6 min.

1	B150cm_l - A75_r	b a	Repetitions = 3
2	A150cm_l - B150_r	b a	Repetitions = 2
3	B75cm_l - A75_r	b A	Repetitions = 2
4	A75cm_l - B75_r	b A	Repetitions = 2
5	B75cm_l - A150_r	b A	Repetitions = 4
6	B150cm_l - A150_r	B A	Repetitions = 2
7	A150cm_l - B75_r	B A	Repetitions = 2
8	B75cm_l - A75_r	b a	Repetitions = 4
9	A150cm_l - B150_r	A b	Repetitions = 2
10	A75cm_l - B150_r	A b	Repetitions = 2

Total number of repetitions is 25.

	CD	CD	Image	Reps
1	YY	YN		3
2	NY	NY		2
3	YN	YY		2
4	NN	NY		2
5	YN	YN		4
6	YN	YN		2
7	NN	NY		2
8	YN	YN		4
9	YN	YY		2
10	YY	YY	Y	2

Tot	7 3	7 6	1	25
-----	-----	-----	---	----

Subject 2 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 8 min.

1	A150cm_l - B75_r	b A	Repetitions = 2
2	B150cm_l - A150_r	b A	Repetitions = 3
3	B75cm_l - A75_r	a B	Repetitions = 3
4	B150cm_l - A75_r	b A	Repetitions = 3
5	A75cm_l - B150_r	b a	Repetitions = 2
6	B75cm_l - A150_r	B A	Repetitions = 3
7	B75cm_l - A75_r	b A	Repetitions = 2
8	B150cm_l - A75_r	b A	Repetitions = 2
9	A150cm_l - B150_r	b A	Repetitions = 2
10	A75cm_l - B75_r	a B	Repetitions = 2

Total number of repetitions is 24.

	CD	CD	Image	Reps
1	NY	NY		2
2	YY	YN		3
3	NN	NY		3
4	YY	YY	Y	3
5	NN	NY		2
6	YY	YN		3
7	YN	YY		2
8	YY	YY	Y	2
9	NY	NN		2
10	YN	YY		2
Tot	6 6	6 7	2	24

Subject 2 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 6 min.

1	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 2
2	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 2
3	A75cm_l - B75_r - cartoon depth	A B	Repetitions = 2
4	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 1
5	B75cm_l - A75_r - cartoon depth	A B	Repetitions = 3
6	B75cm_l - A150_r - cartoon depth	A b	Repetitions = 3
7	B75cm_l - A75_r - cartoon depth	A B	Repetitions = 3
8	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 2
9	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 1
10	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2

Total number of repetitions is 21.

	CD	CD	Image	Reps
1	YY	YY	Y	2
2	YY	YY	Y	2
3	YY	YY	Y	2
4	NY	NY		1
5	NY	NY		3
6	NY	NY		3
7	NY	NY		3
8	YY	YY	Y	2
9	NY	NY		1
10	YY	YY	Y	2
<hr/>				
Tot	5 10	5 10	5	21

Subject 3 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 8 min.

1	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 4
2	A75cm_l - B150_r - cartoon depth	a B	Repetitions = 10
3	A75cm_l - B75_r - cartoon depth	B A	Repetitions = 8
4	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 3
5	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 4
6	B150cm_l - A75_r - cartoon depth	b A	Repetitions = 2
7	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 3
8	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 2
9	B75cm_l - A75_r - cartoon depth	b A	Repetitions = 3
10	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 3

Total number of repetitions is 42.

	CD	CD	Image	Reps
1	YY	YY	Y	4
2	YN	YN		10
3	NY	NY		8
4	YY	YY	Y	3
5	YY	YY	Y	4
6	YY	YY	Y	2
7	NY	NY		3
8	NY	NY		2
9	YN	YY		3
10	YY	YY	Y	3
Tot	7 8	7 9	5	42

Subject 3 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 9 min.

1	A75cm_l - B75_r	b a	Repetitions = 6
2	A150cm_l - B75_r	a B	Repetitions = 6
3	A150cm_l - B150_r	a b	Repetitions = 4
4	B150cm_l - A75_r	b A	Repetitions = 6
5	B75cm_l - A75_r	B A	Repetitions = 2
6	B75cm_l - A150_r	B a	Repetitions = 4
7	A150cm_l - B150_r	B a	Repetitions = 5
8	B150cm_l - A150_r	b a	Repetitions = 6
9	B150cm_l - A75_r	B A	Repetitions = 7
10	A75cm_l - B150_r	A b	Repetitions = 3

Total number of repetitions is 49.

	CD	CD	Image	Reps
1	NN	NN		6
2	YY	YY	Y	6
3	YY	YY	Y	4
4	YY	YY	Y	6
5	YY	YY	Y	2
6	YY	YY	Y	4
7	NN	NY		5
8	YY	YY	Y	6
9	YN	YY		7
10	YY	YY	Y	3

Tot	8 7	8 9	7	49
-----	-----	-----	---	----

Subject 4 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 3 min.

1	A75cm_l - B75_r	A B	Repetitions = 2
2	B75cm_l - A150_r	B a	Repetitions = 1
3	B150cm_l - A75_r	b a	Repetitions = 2
4	A150cm_l - B75_r	a B	Repetitions = 2
5	B150cm_l - A150_r	b a	Repetitions = 2
6	A150cm_l - B75_r	a B	Repetitions = 1
7	B75cm_l - A75_r	B a	Repetitions = 1
8	A75cm_l - B150_r	A b	Repetitions = 1
9	B150cm_l - A75_r	b a	Repetitions = 2
10	A150cm_l - B150_r	a b	Repetitions = 2

Total number of repetitions is 16.

	CD	CD	Image	Reps
1	YY	YY	Y	2
2	YY	YY	Y	1
3	YY	YN		2
4	YY	YY	Y	2
5	YY	YY	Y	2
6	YY	YY	Y	1
7	YY	YN		1
8	YY	YY	Y	1
9	YY	YN		2
10	YY	YY	Y	2
Tot	10 10	10 7	7	16

Subject 4 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 3 min.

1	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 1
2	A150cm_l - B75_r - cartoon depth	a B	Repetitions = 1
3	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 1
4	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 2
5	B150cm_l - A75_r - cartoon depth	b A	Repetitions = 1
6	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 2
7	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2
8	A75cm_l - B75_r - cartoon depth	A B	Repetitions = 1
9	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 1
10	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 1

Total number of repetitions is 13.

	CD	CD	Image	Reps
1	YY	YY	Y	1
2	YY	YY	Y	1
3	YY	YY	Y	1
4	YY	YY	Y	2
5	YY	YY	Y	1
6	NY	NY		2
7	YY	YY	Y	2
8	YY	YY	Y	1
9	YY	YY	Y	1
10	YY	YY	Y	1

Tot	9 10	9 10	8	13
-----	------	------	---	----

Subject 5 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 4 min.

1	B75cm_l - A75_r - cartoon depth	A B	Repetitions = 3
2	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 2
3	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 2
4	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2
5	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 2
6	A150cm_l - B75_r - cartoon depth	B a	Repetitions = 3
7	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 2
8	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 3
9	A75cm_l - B75_r - cartoon depth	A B	Repetitions = 2
10	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 2

Total number of repetitions is 23.

	CD	CD	Image	Reps
1	NY	NY		3
2	NY	NY		2
3	YY	YY	Y	2
4	YY	YY	Y	2
5	NY	NY		2
6	NN	NN		3
7	YY	YY	Y	2
8	YY	YY	Y	3
9	YY	YY	Y	2
10	YY	YY	Y	2
Tot	6 9	6 9	6	23 reps

Subject 5 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 4 min.

1	B150cm_l - A150_r	b a	Repetitions = 4
2	A75cm_l - B75_r	b A	Repetitions = 3
3	B150cm_l - A75_r	a B	Repetitions = 3
4	A75cm_l - B150_r	b a	Repetitions = 3
5	A150cm_l - B75_r	b A	Repetitions = 2
6	A150cm_l - B75_r	a B	Repetitions = 2
7	B75cm_l - A150_r	b a	Repetitions = 4
8	B75cm_l - A75_r	b A	Repetitions = 4
9	A150cm_l - B150_r	a B	Repetitions = 2
10	B150cm_l - A75_r	a b	Repetitions = 1

Total number of repetitions is 28.

	CD	CD	Image	Reps
1	YY	YY	Y	4
2	NN	NY		3
3	NY	NY		3
4	NN	NY		3
5	NY	NY		2
6	YY	YY	Y	2
7	YN	YY		4
8	YN	YY		4
9	YY	YN		2
10	NY	NN		1

Tot	5 6	5 8	2	28
-----	-----	-----	---	----

Subject 6 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 6 min.

1	A150cm_l - B150_r	a B	Repetitions = 4
2	A150cm_l - B150_r	a B	Repetitions = 2
3	A150cm_l - B75_r	a B	Repetitions = 3
4	A150cm_l - B150_r	A B	Repetitions = 4
5	A75cm_l - B75_r	A B	Repetitions = 2
6	B75cm_l - A150_r	a B	Repetitions = 2
7	B150cm_l - A75_r	b A	Repetitions = 3
8	A75cm_l - B150_r	A B	Repetitions = 3
9	B150cm_l - A150_r	a B	Repetitions = 2
10	B75cm_l - A75_r	a B	Repetitions = 2

Total number of repetitions is 27.

	CD	CD	Image	Reps
1	YY	YN		4
2	YY	YN		2
3	YY	YY	Y	3
4	YN	YN		4
5	YY	YY	Y	2
6	NN	NN		2
7	YY	YY	Y	3
8	YY	YN		3
9	NY	NN		2
10	NN	NY		2

Tot	7 7	7 4	3	27
-----	-----	-----	---	----

Subject 6 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 5 min.

1	A75cm_l - B75_r - cartoon depth	a b	Repetitions = 2
2	A150cm_l - B75_r - cartoon depth	a B	Repetitions = 2
3	B75cm_l - A150_r - cartoon depth	b a	Repetitions = 3
4	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2
5	B75cm_l - A150_r - cartoon depth	B A	Repetitions = 3
6	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 3
7	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 1
8	A75cm_l - B150_r - cartoon depth	a b	Repetitions = 3
9	B150cm_l - A150_r - cartoon depth	a B	Repetitions = 3
10	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 2

Total number of repetitions is 24.

	CD	CD	Image	Reps
1	YN	YN		2
2	YY	YY	Y	2
3	YN	YY		3
4	YY	YY	Y	2
5	YY	YN		3
6	YY	YY	Y	3
7	NY	NY		1
8	YN	YY		3
9	NY	NN		3
10	YY	YY	Y	2
Tot	8 7	8 7	4	24

Subject 7 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 6 min.

1	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 4
2	A150cm_l - B150_r - cartoon depth	b a	Repetitions = 2
3	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 3
4	B75cm_l - A150_r - cartoon depth	A B	Repetitions = 3
5	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 2
6	B75cm_l - A75_r - cartoon depth	A B	Repetitions = 4
7	A75cm_l - B75_r - cartoon depth	a B	Repetitions = 3
8	B150cm_l - A150_r - cartoon depth	a B	Repetitions = 4
9	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 3
10	B150cm_l - A75_r - cartoon depth	b A	Repetitions = 3

Total number of repetitions is 31.

	CD	CD	Image	Reps
1	NY	NY		4
2	NY	NN		2
3	YY	YY	Y	3
4	NY	NN		3
5	NY	NY		2
6	NY	NY		4
7	YN	YY		3
8	NY	NN		4
9	YY	YY	Y	3
10	YY	YY	Y	3
Tot	4 9	4 7	3	31

Subject 7 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 6 min.

1	B150cm_l - A75_r	b A	Repetitions = 2
2	B75cm_l - A75_r	A b	Repetitions = 3
3	A75cm_l - B150_r	a B	Repetitions = 3
4	A75cm_l - B150_r	A B	Repetitions = 3
5	B75cm_l - A150_r	a b	Repetitions = 4
6	A150cm_l - B150_r	A b	Repetitions = 3
7	A75cm_l - B150_r	a b	Repetitions = 3
8	A75cm_l - B75_r	B a	Repetitions = 3
9	A150cm_l - B75_r	B A	Repetitions = 3
10	B150cm_l - A150_r	A b	Repetitions = 4

Total number of repetitions is 31.

	CD	CD	Image	Reps
1	YY	YY	Y	2
2	NY	NN		3
3	YN	YN		3
4	YY	YN		3
5	NN	NY		4
6	YN	YY		3
7	YN	YY		3
8	NY	NN		3
9	NN	NY		3
10	NN	NY		4
Tot	5 4	5 6	1	31

Subject 8 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 5 min.

1	A150cm_l - B150_r	b A	Repetitions = 3
2	B150cm_l - A150_r	a b	Repetitions = 1
3	B150cm_l - A150_r	B a	Repetitions = 1
4	A150cm_l - B150_r	b a	Repetitions = 1
5	B75cm_l - A75_r	A b	Repetitions = 1
6	A75cm_l - B75_r	a b	Repetitions = 1
7	A150cm_l - B75_r	a B	Repetitions = 2
8	B75cm_l - A150_r	A b	Repetitions = 2
9	B150cm_l - A75_r	B a	Repetitions = 1
10	A75cm_l - B150_r	b a	Repetitions = 1

Total number of repetitions is 14.

	CD	CD	Image	Reps
1	NY	NN		3
2	NY	NY		1
3	YN	YY		1
4	NY	NY		1
5	NY	NN		1
6	YN	YN		1
7	YY	YY	Y	2
8	NY	NY		2
9	YN	YN		1
10	NN	NY		1

Tot	4 6	4 6	1	14 reps
-----	-----	-----	---	---------

Subject 8 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 4 min.

1	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 1
2	B75cm_l - A150_r - cartoon depth	b a	Repetitions = 1
3	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 1
4	A75cm_l - B75_r - cartoon depth	B a	Repetitions = 1
5	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2
6	B75cm_l - A75_r - cartoon depth	B a	Repetitions = 1
7	A75cm_l - B150_r - cartoon depth	a B	Repetitions = 1
8	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 1
9	A150cm_l - B75_r - cartoon depth	a B	Repetitions = 1
10	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 1

Total number of repetitions is 11.

	CD	CD	Image	Reps
1	NY	NY		1
2	YN	YY		1
3	YY	YY	Y	1
4	NY	NN		1
5	YY	YY	Y	2
6	YY	YN		1
7	YN	YN		1
8	YY	YY	Y	1
9	YY	YY	Y	1
10	YY	YY	Y	1
Tot	8 8	8 7	5	11 reps

Subject 9 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 5 min.

1	A75cm_l - B75_r - cartoon depth	A b	Repetitions = 1
2	B75cm_l - A150_r - cartoon depth	A b	Repetitions = 3
3	B75cm_l - A75_r - cartoon depth	a B	Repetitions = 1
4	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 1
5	B150cm_l - A75_r - cartoon depth	A B	Repetitions = 1
6	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 1
7	A150cm_l - B75_r - cartoon depth	a B	Repetitions = 1
8	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 1
9	A75cm_l - B75_r - cartoon depth	B a	Repetitions = 1
10	B75cm_l - A150_r - cartoon depth	a B	Repetitions = 1

Total number of repetitions is 12.

	CD	CD	Image	Reps
1	YY	YN		1
2	NY	NY		3
3	NN	NY		1
4	YY	YY	Y	1
5	NN	NY		1
6	NY	NY		1
7	YY	YY	Y	1
8	YY	YY	Y	1
9	NY	NN		1
10	NN	NN		1

Tot	4 7	4 7	3	12 reps
-----	-----	-----	---	---------

Subject 9 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 5 min.

1	A75cm_l - B150_r	a b	Repetitions = 1
2	B75cm_l - A150_r	B a	Repetitions = 1
3	B150cm_l - A150_r	a B	Repetitions = 2
4	A75cm_l - B75_r	a B	Repetitions = 1
5	B150cm_l - A75_r	b A	Repetitions = 2
6	B75cm_l - A75_r	b A	Repetitions = 2
7	A150cm_l - B75_r	B a	Repetitions = 1
8	A150cm_l - B150_r	b A	Repetitions = 1
9	A75cm_l - B75_r	b A	Repetitions = 1
10	B75cm_l - A150_r	A b	Repetitions = 1

Total number of repetitions is 13.

	CD	CD	Image	Reps
1	YN	YY		1
2	YY	YY	Y	1
3	NY	NN		2
4	YN	YY		1
5	YY	YY	Y	2
6	YN	YY		2
7	NN	NN		1
8	NY	NN		1
9	NN	NY		1
10	NY	NY		1

Tot	5 5	5 7	2	13 reps
-----	-----	-----	---	---------

Subject 10 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 6 min.

1	A75cm_l - B75_r	a b	Repetitions = 2
2	B75cm_l - A75_r	A B	Repetitions = 3
3	A75cm_l - B150_r	A B	Repetitions = 2
4	A75cm_l - B150_r	a B	Repetitions = 5
5	B75cm_l - A150_r	a b	Repetitions = 2
6	B150cm_l - A75_r	A B	Repetitions = 2
7	B150cm_l - A150_r	B A	Repetitions = 2
8	A150cm_l - B75_r	a b	Repetitions = 2
9	A150cm_l - B150_r	a b	Repetitions = 2
10	A75cm_l - B75_r	b a	Repetitions = 2

Total number of repetitions is 24.

	CD	CD	Image	Reps
1	YN	YN		2
2	NY	NY		3
3	YY	YN		2
4	YN	YN		5
5	NN	NY		2
6	NN	NY		2
7	YN	YN		2
8	YY	YN		2
9	YY	YY	Y	2
10	NN	NN		2

Tot	6 4	6 4	1	24 reps
-----	-----	-----	---	---------

Subject 10 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 5 min.

1	A75cm_l - B75_r - cartoon depth	A B	Repetitions = 2
2	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 2
3	B75cm_l - A75_r - cartoon depth	a B	Repetitions = 2
4	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 2
5	A75cm_l - B150_r - cartoon depth	b a	Repetitions = 2
6	B75cm_l - A150_r - cartoon depth	b a	Repetitions = 3
7	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 1
8	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 1
9	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 1
10	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 2

Total number of repetitions is 18.

	CD	CD	Image	Reps
1	YY	YY	Y	2
2	NY	NY		2
3	NN	NY		2
4	NY	NY		2
5	NN	NY		2
6	YN	YY		3
7	NY	NY		1
8	YY	YY	Y	1
9	YY	YY	Y	1
10	YY	YY	Y	2
<hr/>				
Tot	5 7	5 10	4	18 reps

Subject 11 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 5 min.

1	B150cm_1 - A75_r - cartoon depth	b A	Repetitions = 2
2	A150cm_1 - B75_r - cartoon depth	b A	Repetitions = 2
3	B150cm_1 - A75_r - cartoon depth	a B	Repetitions = 2
4	B150cm_1 - A150_r - cartoon depth	a b	Repetitions = 1
5	A75cm_1 - B150_r - cartoon depth	A B	Repetitions = 2
6	B75cm_1 - A75_r - cartoon depth	B A	Repetitions = 2
7	A150cm_1 - B150_r - cartoon depth	a b	Repetitions = 1
8	A75cm_1 - B75_r - cartoon depth	A B	Repetitions = 2
9	B150cm_1 - A75_r - cartoon depth	a B	Repetitions = 1
10	B75cm_1 - A150_r - cartoon depth	B a	Repetitions = 2

Total number of repetitions is 17.

	CD	CD	Image	Reps
1	YY	YY	Y	2
2	NY	NY		2
3	NY	NY		2
4	NY	NY		1
5	YY	YN		2
6	YY	YY	Y	2
7	YY	YY	Y	1
8	YY	YY	Y	2
9	NY	NY		1
10	YY	YY	Y	2

Tot	6 10	6 9	5	17 reps
-----	------	-----	---	---------

Subject 11 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 5 min.

1	B150cm_l - A150_r	a B	Repetitions = 2
2	B75cm_l - A150_r	A B	Repetitions = 1
3	A150cm_l - B150_r	a B	Repetitions = 2
4	A75cm_l - B150_r	A B	Repetitions = 2
5	B75cm_l - A150_r	B A	Repetitions = 3
6	B75cm_l - A75_r	a B	Repetitions = 1
7	B150cm_l - A75_r	A b	Repetitions = 2
8	A75cm_l - B75_r	a B	Repetitions = 2
9	A150cm_l - B150_r	b A	Repetitions = 1
10	A150cm_l - B75_r	a b	Repetitions = 2

Total number of repetitions is 18.

	CD	CD	Image	Reps
1	NY	NN		2
2	NY	NN		1
3	YY	YN		2
4	YY	YN		2
5	YY	YN		3
6	NN	NY		1
7	NN	NN		2
8	YN	YY		2
9	NY	NN		1
10	YY	YN		2
Tot	5 7	5 2	0	18 reps

Subject 12 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 8 min.

1	A150cm_l - B75_r	a B	Repetitions = 4
2	B150cm_l - A150_r	B A	Repetitions = 3
3	A150cm_l - B150_r	A B	Repetitions = 3
4	A150cm_l - B75_r	a B	Repetitions = 4
5	A75cm_l - B150_r	a B	Repetitions = 3
6	B75cm_l - A75_r	A b	Repetitions = 2
7	A75cm_l - B75_r	a B	Repetitions = 3
8	B75cm_l - A150_r	B A	Repetitions = 3
9	B150cm_l - A75_r	a B	Repetitions = 3
10	A75cm_l - B75_r	b A	Repetitions = 3

Total number of repetitions is 31.

	CD	CD	Image	Reps
1	YY	YY	Y	4
2	YN	YN		3
3	YN	YN		3
4	YY	YY	Y	4
5	YN	YN		3
6	NY	NN		2
7	YN	YY		3
8	YY	YN		3
9	NY	NY		3
10	NN	NY		3
Tot	75	75	2	31 reps

Subject 12 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 5 min.

1	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 3
2	B75cm_l - A150_r - cartoon depth	b a	Repetitions = 2
3	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 2
4	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 3
5	A75cm_l - B150_r - cartoon depth	b a	Repetitions = 2
6	A75cm_l - B75_r - cartoon depth	b a	Repetitions = 2
7	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 3
8	B75cm_l - A75_r - cartoon depth	b A	Repetitions = 3
9	A150cm_l - B150_r - cartoon depth	b a	Repetitions = 3
10	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2

Total number of repetitions is 25.

	CD	CD	Image	Reps
1	NY	NY		3
2	YN	YY		2
3	NY	NY		2
4	NY	NY		3
5	NN	NY		2
6	NN	NN		2
7	NY	NY		3
8	YN	YY		3
9	NY	NY		3
10	YY	YY	Y	2
<hr/>				
Tot	3 6	3 9	1	25 reps

Subject 13 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 4 min.

1	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 1
2	B75cm_l - A150_r - cartoon depth	a b	Repetitions = 2
3	B75cm_l - A75_r - cartoon depth	A B	Repetitions = 2
4	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2
5	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 2
6	B150cm_l - A150_r - cartoon depth	b A	Repetitions = 3
7	B150cm_l - A75_r - cartoon depth	A B	Repetitions = 2
8	B150cm_l - A150_r - cartoon depth	a B	Repetitions = 1
9	A75cm_l - B75_r - cartoon depth	A B	Repetitions = 1
10	A75cm_l - B150_r - cartoon depth	B A	Repetitions = 2

Total number of repetitions is 18.

	CD	CD	Image	Reps
1	NY	NY		1
2	NN	NY		2
3	NY	NY		2
4	YY	YY	Y	2
5	NY	NY		2
6	YN	YN		3
7	NN	NY		2
8	NY	NN		1
9	YY	YY	Y	1
10	NY	NN		2
<hr/>				
Tot	3 7	3 7	2	18 reps

Subject 13 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 5 min.

1	A75cm_1 - B150_r	B A	Repetitions = 1
2	B75cm_1 - A150_r	a B	Repetitions = 2
3	B75cm_1 - A75_r	b A	Repetitions = 2
4	B75cm_1 - A75_r	a B	Repetitions = 2
5	A150cm_1 - B75_r	a B	Repetitions = 3
6	A150cm_1 - B150_r	b A	Repetitions = 1
7	B150cm_1 - A150_r	b A	Repetitions = 3
8	A75cm_1 - B150_r	a B	Repetitions = 1
9	B150cm_1 - A75_r	b A	Repetitions = 1
10	A75cm_1 - B75_r	a B	Repetitions = 1

Total number of repetitions is 17.

	CD	CD	Image	Reps
1	NY	NN		1
2	NN	NN		2
3	YN	YY		2
4	NN	NY		2
5	YY	YY	Y	3
6	NY	NN		1
7	YY	YN		3
8	YN	YN		1
9	YY	YY	Y	1
10	YN	YY		1

Tot	6 5	6 5	2	17 reps
-----	-----	-----	---	---------

Subject 14 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 7 min.

1	B150cm_l - A150_r	A b	Repetitions = 4
2	B150cm_l - A75_r	a B	Repetitions = 3
3	A75cm_l - B150_r	a B	Repetitions = 4
4	A150cm_l - B75_r	A b	Repetitions = 2
5	A75cm_l - B150_r	b A	Repetitions = 2
6	B75cm_l - A75_r	b A	Repetitions = 1
7	A75cm_l - B75_r	b A	Repetitions = 2
8	B150cm_l - A75_r	b A	Repetitions = 2
9	B75cm_l - A150_r	A b	Repetitions = 2
10	A150cm_l - B150_r	a B	Repetitions = 3

Total number of repetitions is 25.

	CD	CD	Image	Reps
1	NN	NY		4
2	NY	NY		3
3	YN	YN		4
4	YN	YN		2
5	NN	NN		2
6	YN	YY		1
7	NN	NY		2
8	YY	YY	Y	2
9	NY	NY		2
10	YY	YN		3

Tot	5 4	5 6	1	25 reps
-----	-----	-----	---	---------

Subject 14 – Test 2

2 min practice. Cartoon depth images. 2.1 sec scan. Time for test = 6 min.

1	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 2
2	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 2
3	B75cm_l - A75_r - cartoon depth	A B	Repetitions = 4
4	B75cm_l - A150_r - cartoon depth	a b	Repetitions = 2
5	A75cm_l - B150_r - cartoon depth	a b	Repetitions = 1
6	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 1
7	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 2
8	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 3
9	A75cm_l - B150_r - cartoon depth	a B	Repetitions = 4
10	A75cm_l - B75_r - cartoon depth	B A	Repetitions = 4

Total number of repetitions is 25.

	CD	CD	Image	Reps
1	NY	NY		2
2	NY	NY		2
3	NY	NY		4
4	NN	NY		2
5	YN	YY		1
6	YY	YY	Y	1
7	NY	NY		2
8	YY	YY	Y	3
9	YN	YN		4
10	NY	NY		4
<hr/>				
Tot	4 7	4 9	2	25 reps

Subject 15 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 7 min.

1	B150cm_l - A75_r - cartoon depth	b A	Repetitions = 3
2	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 4
3	B150cm_l - A150_r - cartoon depth	a B	Repetitions = 4
4	A150cm_l - B75_r - cartoon depth	B a	Repetitions = 6
5	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 9
6	A150cm_l - B150_r - cartoon depth	b a	Repetitions = 2
7	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 3
8	B75cm_l - A150_r - cartoon depth	b a	Repetitions = 6
9	A75cm_l - B75_r - cartoon depth	b A	Repetitions = 4
10	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 2

Total number of repetitions is 43.

	CD	CD	Image	Reps
1	YY	YY	Y	3
2	YY	YY	Y	4
3	NY	NN		4
4	NN	NN		6
5	YY	YY	Y	9
6	NY	NY		2
7	NY	NY		3
8	YN	YY		6
9	NN	NY		4
10	YY	YY	Y	2
<hr/>				
Tot	5 7	5 8	4	43

Subject 15 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 6 min.

1	A75cm_l - B150_r	A b	Repetitions = 3
2	B75cm_l - A75_r	B a	Repetitions = 5
3	A75cm_l - B150_r	a B	Repetitions = 3
4	B150cm_l - A75_r	b A	Repetitions = 2
5	A150cm_l - B75_r	b A	Repetitions = 6
6	B150cm_l - A150_r	b A	Repetitions = 4
7	A150cm_l - B75_r	a B	Repetitions = 2
8	A150cm_l - B150_r	a B	Repetitions = 3
9	A75cm_l - B75_r	a B	Repetitions = 3
10	B75cm_l - A150_r	b A	Repetitions = 2

Total number of repetitions is 33.

	CD	CD	Image	Reps
1	YY	YY	Y	3
2	YY	YN		5
3	YN	YN		3
4	YY	YY	Y	2
5	NY	NY		6
6	YY	YN		4
7	YY	YY	Y	2
8	YY	YN		3
9	YN	YY		3
10	YN	YN		2
Tot	9 7	9 5	3	33

Subject 16 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 8 min.

1	B75cm_l - A75_r	a B	Repetitions = 6
2	B150cm_l - A150_r	A B	Repetitions = 6
3	B150cm_l - A150_r	B A	Repetitions = 4
4	A75cm_l - B75_r	a B	Repetitions = 5
5	B75cm_l - A75_r	b A	Repetitions = 4
6	B150cm_l - A75_r	B a	Repetitions = 6
7	A150cm_l - B75_r	a B	Repetitions = 5
8	A75cm_l - B150_r	A B	Repetitions = 4
9	A150cm_l - B150_r	b A	Repetitions = 5
10	B75cm_l - A150_r	a b	Repetitions = 5

Total number of repetitions is 61.

	CD	CD	Image	Reps
1	NN	NY		6
2	NN	NN		6
3	YN	YN		4
4	YN	YY		5
5	YN	YY		4
6	YN	YN		6
7	YY	YY	Y	5
8	YY	YN		4
9	NY	NN		5
10	NN	NY		5
Tot	6 3	6 5	1	50

Subject 16 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 5 min.

1	B150cm_l - A75_r - cartoon depth	b A	Repetitions = 4
2	A150cm_l - B75_r - cartoon depth	a B	Repetitions = 4
3	B75cm_l - A150_r - cartoon depth	a b	Repetitions = 3
4	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 2
5	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 3
6	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 3
7	A75cm_l - B150_r - cartoon depth	a B	Repetitions = 3
8	A75cm_l - B75_r - cartoon depth	B A	Repetitions = 2
9	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 4
10	B75cm_l - A150_r - cartoon depth	A B	Repetitions = 3

Total number of repetitions is 28.

	CD	CD	Image	Reps
1	YY	YY	Y	4
2	YY	YY	Y	4
3	NN	NY		3
4	NY	NY		2
5	YY	YY	Y	3
6	YY	YY	Y	3
7	YN	YN		3
8	NY	NY		2
9	NY	NY		4
10	NY	NN		3
<hr/>				
Tot	5 8	5 8	4	31

Subject 17 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 6 min.

1	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 5
2	A75cm_l - B75_r - cartoon depth	A B	Repetitions = 3
3	A75cm_l - B150_r - cartoon depth	a B	Repetitions = 3
4	A75cm_l - B150_r - cartoon depth	a b	Repetitions = 6
5	B75cm_l - A75_r - cartoon depth	b A	Repetitions = 3
6	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 1
7	B150cm_l - A75_r - cartoon depth	b A	Repetitions = 3
8	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 2
9	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 3
10	B75cm_l - A150_r - cartoon depth	A B	Repetitions = 7

Total number of repetitions is 36.

	CD	CD	Image	Reps
1	YY	YY	Y	5
2	YY	YY	Y	3
3	YN	YN		3
4	YN	YY		6
5	YN	YY		3
6	NY	NY		1
7	YY	YY	Y	3
8	NY	NY		2
9	NY	NY		3
10	NY	NN		7
<hr/>				
Tot	6 7	6 8	3	36

Subject 17 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 8 min.

1	A75cm_1 - B150_r	b A	Repetitions = 4
2	A75cm_1 - B75_r	a B	Repetitions = 4
3	A150cm_1 - B150_r	a B	Repetitions = 6
4	B150cm_1 - A150_r	A B	Repetitions = 4
5	B150cm_1 - A75_r	a B	Repetitions = 6
6	B75cm_1 - A150_r	a b	Repetitions = 4
7	B75cm_1 - A75_r	a B	Repetitions = 4
8	A150cm_1 - B75_r	A B	Repetitions = 5
9	A150cm_1 - B75_r	b A	Repetitions = 6
10	B150cm_1 - A150_r	a B	Repetitions = 10

Total number of repetitions is 53.

	CD	CD	Image	Reps
1	NN	NN		4
2	YN	YY		4
3	YY	YN		6
4	NN	NN		4
5	NY	NY		6
6	NN	NY		4
7	NN	NY		4
8	YN	YY		5
9	NY	NY		6
10	NY	NN		10
Tot	3 4	3 6	0	53

Subject 18 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 8 min.

1	A150cm_l - B75_r	a B	Repetitions = 3
2	B150cm_l - A150_r	A B	Repetitions = 5
3	B75cm_l - A150_r	A b	Repetitions = 4
4	B75cm_l - A75_r	A b	Repetitions = 6
5	B150cm_l - A75_r	b A	Repetitions = 5
6	B150cm_l - A150_r	a B	Repetitions = 5
7	A150cm_l - B150_r	A b	Repetitions = 5
8	A75cm_l - B150_r	b A	Repetitions = 4
9	A75cm_l - B75_r	a B	Repetitions = 6
10	A75cm_l - B75_r	a b	Repetitions = 5

Total number of repetitions is 48.

	CD	CD	Image	Reps
1	YY	YY	Y	3
2	NN	NN		5
3	NY	NY		4
4	NY	NN		6
5	YY	YY	Y	5
6	NY	NN		5
7	YN	YY		5
8	NN	NN		4
9	YN	YY		6
10	YN	YN		5
<hr/>				
Tot	5 5	5 5	2	48

Subject 18 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 4 min.

1	A75cm_l - B75_r - cartoon depth	b A	Repetitions = 3
2	A150cm_l - B150_r - cartoon depth	a B	Repetitions = 2
3	B75cm_l - A150_r - cartoon depth	A b	Repetitions = 2
4	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 3
5	B75cm_l - A75_r - cartoon depth	a B	Repetitions = 3
6	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 2
7	A75cm_l - B150_r - cartoon depth	a b	Repetitions = 2
8	B150cm_l - A150_r - cartoon depth	b a	Repetitions = 3
9	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 2
10	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 4

Total number of repetitions is 26.

	CD	CD	Image	Reps
1	NN	NY		3
2	YY	YN		2
3	NY	NY		2
4	NY	NY		3
5	NN	NY		3
6	NY	NY		2
7	YN	YY		2
8	YY	YY	Y	3
9	NY	NY		2
10	NY	NY		4
<hr/>				
Tot	3 7	3 9	1	26

Subject 19 – Test 1

2 min practice. Cartoon depth images. 2 second scan. Time for test = 6 min.

1	B75cm_l - A150_r - cartoon depth	b a	Repetitions = 4
2	A75cm_l - B75_r - cartoon depth	a B	Repetitions = 4
3	A75cm_l - B150_r - cartoon depth	b a	Repetitions = 3
4	A75cm_l - B150_r - cartoon depth	a B	Repetitions = 4
5	A75cm_l - B150_r - cartoon depth	a b	Repetitions = 3
6	B75cm_l - A75_r - cartoon depth	A B	Repetitions = 4
7	A150cm_l - B150_r - cartoon depth	a b	Repetitions = 2
8	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 2
9	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 3
10	A150cm_l - B75_r - cartoon depth	b A	Repetitions = 3

Total number of repetitions is 32.

	CD	CD	Image	Reps
1	YN	YY		4
2	YN	YY		4
3	NN	NY		3
4	YN	YN		4
5	YN	YY		3
6	NY	NY		4
7	YY	YY	Y	2
8	NY	NY		2
9	NY	NY		3
10	NY	NY		3
<hr/>				
Tot	5 5	5 9	1	32

Subject 19 – Test 2

2 min practice. Plain images. 2 second scan. Time for test = 5 min.

1	A75cm_l - B75_r	B A	Repetitions = 4
2	B150cm_l - A150_r	A B	Repetitions = 3
3	B75cm_l - A75_r	b A	Repetitions = 2
4	B150cm_l - A75_r	a B	Repetitions = 4
5	A150cm_l - B150_r	b A	Repetitions = 1
6	A150cm_l - B75_r	a B	Repetitions = 2
7	B75cm_l - A150_r	b A	Repetitions = 3
8	A75cm_l - B150_r	a B	Repetitions = 4
9	B75cm_l - A75_r	b A	Repetitions = 3
10	B150cm_l - A150_r	a B	Repetitions = 3

Total number of repetitions is 29.

	CD	CD	Image	Reps
1	NY	NY		4
2	NN	NN		3
3	YN	YY		2
4	NY	NY		4
5	NY	NN		1
6	YY	YY	Y	2
7	YN	YN		3
8	YN	YN		4
9	YN	YY		3
10	NY	NN		3
<hr/>				
Tot	5 5	5 5	1	29

Subject 20 – Test 1

2 min practice. Plain images. 2 second scan. Time for test = 5 min.

1	B150cm_l - A150_r	b A	Repetitions = 4
2	A75cm_l - B150_r	A B	Repetitions = 3
3	A150cm_l - B150_r	A b	Repetitions = 3
4	B75cm_l - A75_r	B a	Repetitions = 3
5	A150cm_l - B75_r	A b	Repetitions = 4
6	A150cm_l - B150_r	b a	Repetitions = 3
7	B150cm_l - A150_r	a b	Repetitions = 4
8	B150cm_l - A75_r	b A	Repetitions = 2
9	B75cm_l - A150_r	a b	Repetitions = 3
10	A75cm_l - B75_r	A B	Repetitions = 3

Total number of repetitions is 32.

	CD	CD	Image	Reps
1	YY	YN		4
2	YY	YN		3
3	YN	YY		3
4	YY	YN		3
5	YN	YN		4
6	NY	NY		3
7	NY	NY		4
8	YY	YY	Y	2
9	NN	NY		3
10	YY	YY	Y	3

Tot	77	76	2	32
-----	----	----	---	----

Subject 20 – Test 2

2 min practice. Cartoon depth images. 2 second scan. Time for test = 4 min.

1	B75cm_l - A150_r - cartoon depth	B a	Repetitions = 2
2	B150cm_l - A75_r - cartoon depth	a B	Repetitions = 4
3	A150cm_l - B75_r - cartoon depth	a B	Repetitions = 1
4	A75cm_l - B150_r - cartoon depth	A b	Repetitions = 3
5	A75cm_l - B75_r - cartoon depth	A b	Repetitions = 2
6	B75cm_l - A75_r - cartoon depth	B A	Repetitions = 3
7	B150cm_l - A150_r - cartoon depth	a b	Repetitions = 3
8	A75cm_l - B150_r - cartoon depth	A B	Repetitions = 2
9	A75cm_l - B150_r - cartoon depth	A B	Repetitions = 2
10	A150cm_l - B150_r - cartoon depth	b a	Repetitions = 3

Total number of repetitions is 25.

	CD	CD	Image	Reps
1	YY	YY	Y	2
2	NY	NY		4
3	YY	YY	Y	1
4	YY	YY	Y	3
5	YY	YN		2
6	YY	YY	Y	3
7	NY	NY		3
8	YY	YN		2
9	YY	YN		2
10	NY	NY		3
Tot	7 10	7 7	4	25

Appendix 3. Hardware & software considerations

In this chapter various aspects of the hardware and software used during the research are considered. A great deal of time and effort was spent in setting up suitable equipment to capture stereo images that could be fed directly to a standard computer. Once accomplished an even greater length of time was dedicated to the development of suitable software programmes that could be used to process the live video feed into stereo depth maps and to finally perform the conversion to optophonic sound.

A3.1. Hardware

During the three years of research the equipment used has varied greatly due to the rapid advances in technology and computers. Originally, before considering stereovision for use with a blind mobility aid a single camera was used in conjunction with a Silicon Graphics O2. This meant that initially all software was written in C++ and ran under a Unix operating system. At the time this computer was far superior to any PC available and supplied a suitable basis for investigating alternative algorithms for use with the different aspects of optophonic mappings, since it was more than capable of processing multiple frames per second. The reasoning behind this scheme was that once a suitable algorithm had been found, it could be optimised for use on a PC.

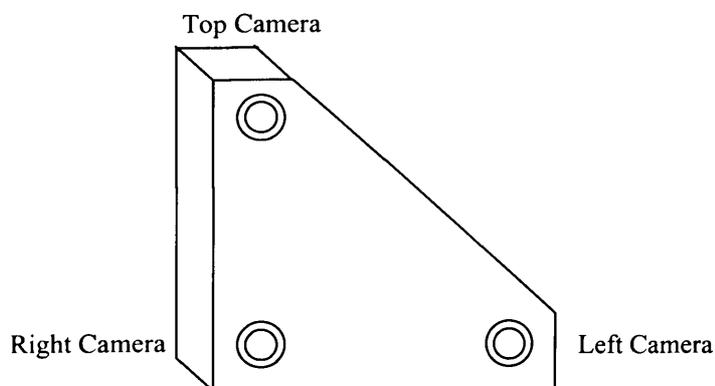


Figure (A3.1) – Diagram of the Triclops camera system.

Once it was realised that an optophonic blind aid could be benefited by depth information as obtained through the use of multiple cameras, work began studying the ways in which a multi-camera system could be created. Three mini-cameras were purchased, the signals from which could be combined into the separate RGB (Red, Green, Blue) of a computer video card and extracted by software. The only problem left to solve was how to synchronise the three signals from the separate cameras so that each captured frames simultaneously. However, it was unnecessary to solve this problem since at this time a three-camera system named Triclops became available. This system, which was purchased from Point Grey Research in Canada [PGR00], consisted of a set of libraries for use in Microsoft Visual C++ with which it is possible to access the images obtained from the Triclops cameras. Therefore, all software had to be rewritten in Visual C++ in order to make use of the Triclops libraries.

The three-camera module, shown in figure (A3.1), consists of three gray-scale, 1/3'' CCD cameras. The device itself has been designed to work with and plug into the RGB input on a Matrox Meteor RGB graphics card fitted into a PC.

A3.2. Software

The software used during the research for the purposes of testing various modifications to the optophonic mapping from scene-to-sound was developed for various tasks using a range of compilers. In the following sections a brief description shall be given of the compilers used and a brief explanation of some of the more important programmes written.

A3.2.1. Software compilers

All algorithms described within this thesis and used during the research were written in Microsoft Visual C++ 6.0 and implemented using the Triclops system and libraries. However, in a number of cases algorithms were first tested with still images in C++ on a powerful Silicon Graphics or in Borland Builder C++ 3.0 on a PC due to its ease

of use and the speed with which it is possible to produce a full working programme with graphical user interface (most programmes in Borland Builder can be created with only a few lines of code). When a suitable algorithm had been developed and tested with still images it was then ported across to Microsoft Visual C++ allowing the use of the Triclops system for real-time image capturing and processing.

Due to the various compilers that have been used to test and develop various algorithms a number of different standalone programmes were created. For instance, as mentioned above, a programme was written in Borland Builder C++ that originally consisted of a simple interface that allowed the user to open stereo images, manipulate them (using simple image processing filters), and then save them to a different file. Once this programme had been fully developed it provided a suitable basis to test new algorithms. Through the simple addition of a new function (that performed the algorithm to be tested) to the main programme any technique could be tested with still images.

Similarly, since the Triclops camera system consisted of control libraries that could only be accessed through Microsoft Visual C++, a programme was developed that could be used as a base for future development. With this programme any algorithms could be simply applied and tested in real-time using images captured directly from the Triclops cameras.

A3.2.2. Test programmes

Out of the many programmes that were written during the research there were three that were used as a basis for future development, from which further modifications could easily be carried out to test new ideas. Two of these programmes were written and compiled in Borland Builder C++ for the purposes of testing various stereo processing algorithms on still images, and secondly for testing volunteers with the scene-to-sound mappings. The final programme, written in Microsoft Visual C++, was designed to test in real-time the various modifications to the optophonic process with images captured directly from the Triclops camera system.

A3.2.2.1. Stereo algorithm programme

The programme created in Borland Builder to assess various stereo algorithms with still images allows the user to search for and load stereo images from disk. When loaded various previously hidden buttons appear which provide the user with a choice of stereo algorithms and image filters. These differing algorithms can be applied in any order and with different combinations of filters. Once processing has ceased, the user is provided with the option to save the images to file simply by clicking on the appropriate button. The various options can be selected by opening the options dialog box from the menu. Within this dialog the choices can be made upon which filters to apply and on the various disparity ranges to use.



Figure (A3.2) – A screenshot of the programme written to test various stereo algorithms and other image processing techniques.

Figure (A3.2) shows a screenshot taken from the main window of the programme and figures (A3.3a)-(A3.3c) demonstrate the options window, from which it is possible to select various image processing options. Although the programme may not be pretty, it does provide a very easy way of testing new algorithms by simply placing another switch or check box within the options dialog box and then linking it to a new function. After the initial programme engine had been created future functions could just access the library of commands already coded making further additions relatively simple.

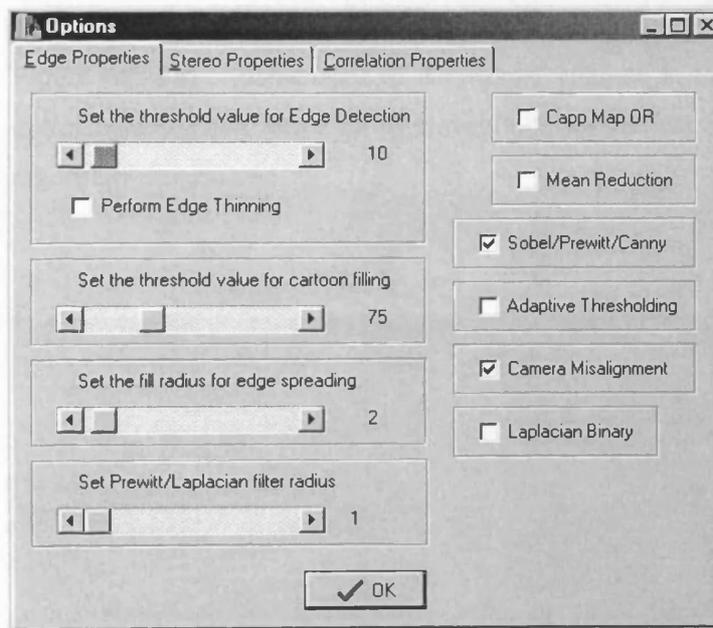


Figure (A3.3a) – Screenshot of the first window in the options dialog box revealing edge and image filter options for the stereo algorithm programme.

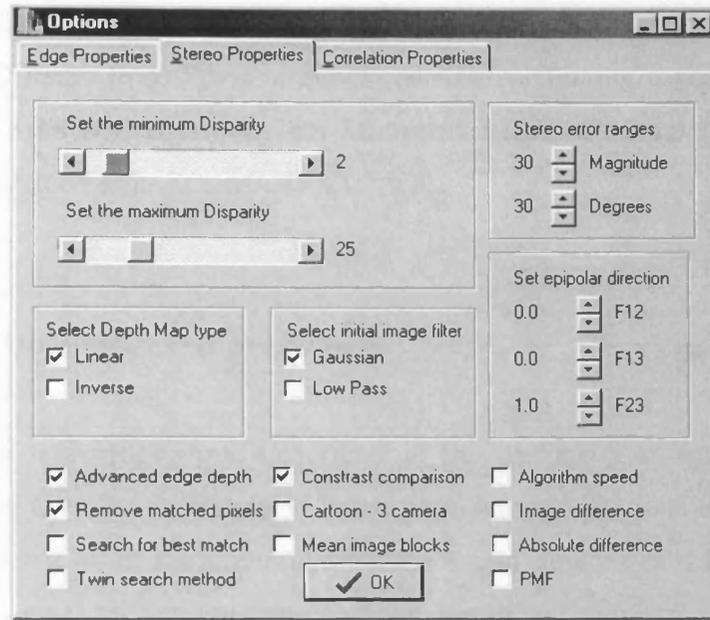


Figure (A3.3b) – Screenshot of the second window in the options dialog box showing stereo options for the stereo algorithm programme.

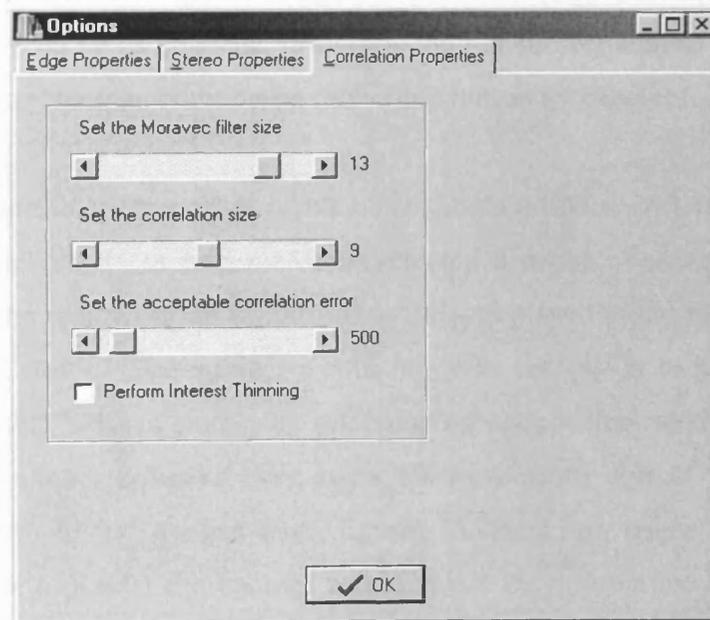


Figure (A3.3c) – Screenshot of the last window in the options dialog box showing correlation properties for the stereo algorithm programme.

After testing algorithms with still images using the programme described (and shown in figure (A3.2)), the techniques or functions used, assuming a positive result was obtained, would be ported across into Microsoft Visual C++ 6.0 by adding to the programme described later in section (A3.2.2.3).

A3.2.2.2. Optophone test programme

The optophone test programme described in this section was again generated in Borland Builder C++ 3 and was used to test various modifications to the optophonic process (whether it be to the processing of the sound, images or the actual mapping from scene-to-sound) by assessing the performance of test subjects.

The programme in its initial form provided numerous options for the conversion of image-to-sound such as the speed of playback, the use of dichotic presentation and other stereo sound effects, as well as the ability to convert video and animations to sound. For the final test programme however, it was thought best that the programme be as simple to operate as possible, providing the test subject with as few situations as possible for error (such as clicking on the wrong button by mistake).

The test programme, a screenshot of which is shown in figure (A3.4a), would ask the user to click on 'start' to begin. Once selected a series of images, although not visible, would be read from an initialisation file and played in the form of sound in a pseudo-random order. The subject would be given the option to play the sound as many times as they wished simply by selecting 'repeat', as depicted in figure (A3.4b), and when the subject believed they knew the location of objects within the scene (such as the 'A' & 'B' images from figures (5.7a)-(5.7p), pages 138-140, and as shown in figure (A3.4d)) the subject would select the appropriate check boxes. A simple graphical representation would be displayed for confirmation (for sighted users), and then a button labelled 'accept' would appear allowing the subject to continue, shown in figure (A3.4c). This process continues until all test images have been displayed, at which time the programme saves the data to an output file for later assessment.

To further reduce the complexity of the test programme, for the user, all options are defined within an initialisation file and are loaded upon execution of the programme itself.

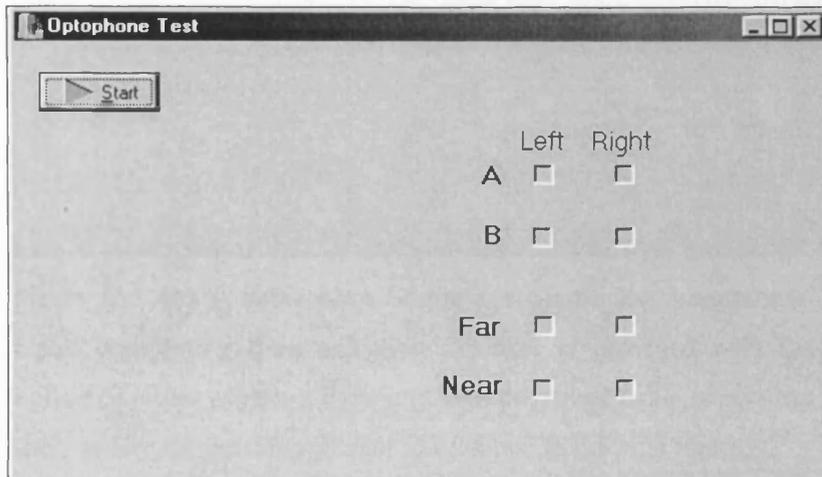


Figure (A3.4a) – The initial screen displayed after executing the optophone test programme. All buttons remain hidden or disabled until the appropriate time. The user must select *Start* to begin.

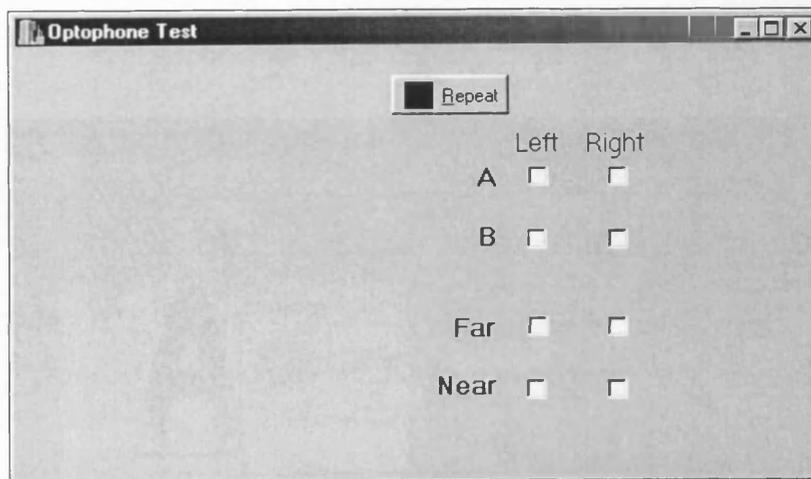


Figure (A3.4b) – Screenshot of the optophone test programme. Upon hearing an optophonic sound the user is provided with the option of repeating the sound (with *Repeat*), or selecting the checkboxes to indicate the location of the objects within the optophonic soundscape (i.e., 'A' Left Near, 'B' Right Far).

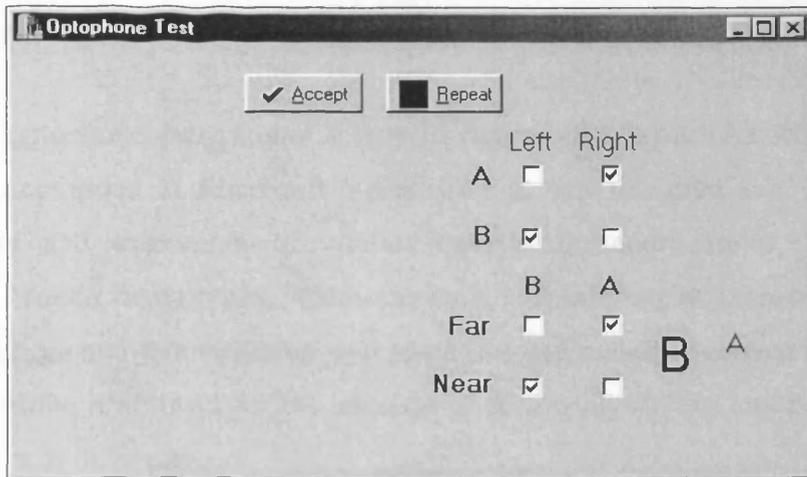


Figure (A3.4c) – Screenshot of the optophone test programme. Upon completing their selection the user is provided with the option of either pressing *Repeat* to hear the sound again to confirm their choice, or pressing *Accept* to continue to the next frame.

The large area on the left of the window, pictured in figures (A3.4a)-(A3.4c), is used during the training period as an image display area as illustrated in figure (A3.4d). This area, as illustrated, is normally hidden and is only displayed when the appropriate settings are used in the initialisation file.

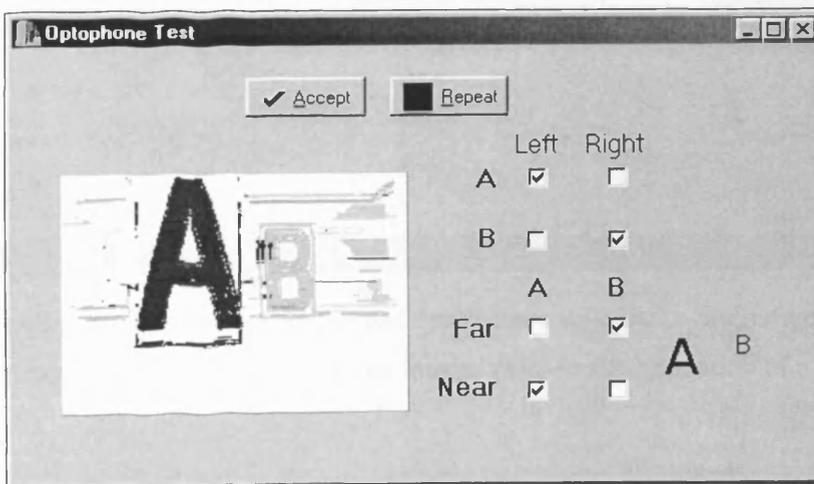


Figure (A3.4d) – Screenshot of the optophone test programme during a training session. This period not only trains the user with the different optophonic sounds, it also provides them with time to become accustomed to the workings of the programme itself.

A3.2.2.3. Real-time stereo optophone programme

The stereo optophone programme shown in figures (A3.5a) & (A3.5b), which was written and compiled in Microsoft Visual C++ 6, was designed as a basis for the development and evaluation of various optophonic modifications, such as the inclusion of stereo depth maps. Once the main software engine (capable of reading both still images and live video, as well as performing real-time conversion to sound) had been written it allowed for the addition of further algorithms and functions with the minimum of difficulty.

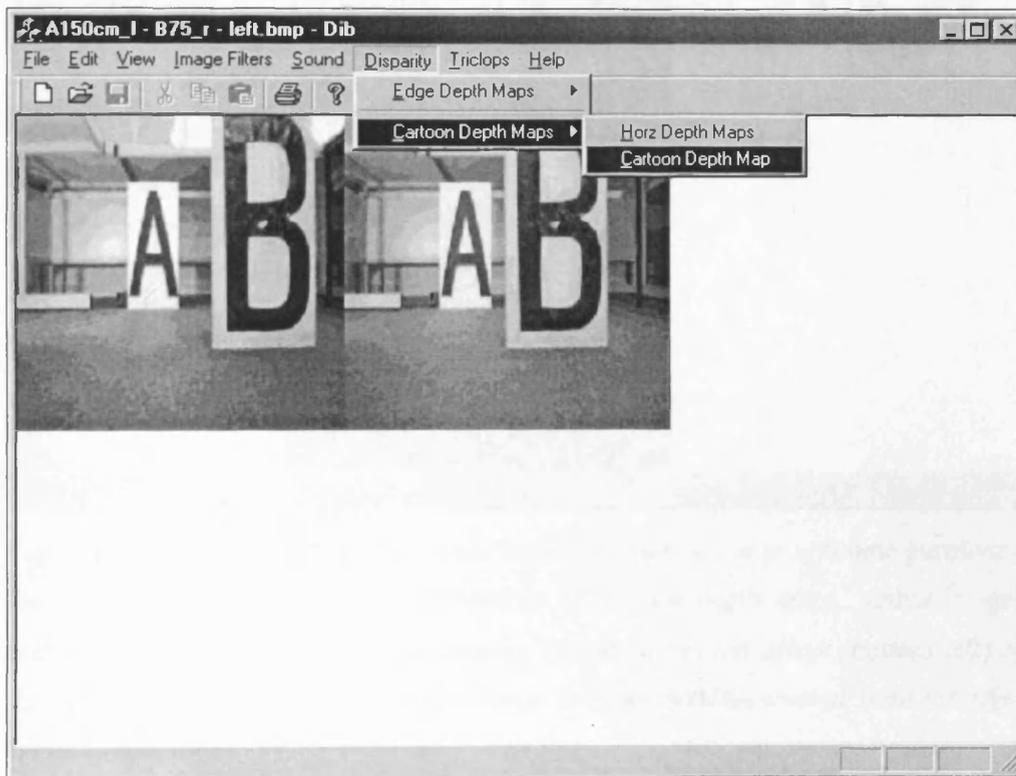


Figure (A3.5a) – Screenshot of the real-time stereo optophone programme. The figure illustrates the capturing of stereo images prior to the generation of a cartoon depth map.

Although this programme was written in Microsoft Visual C++, due to the Triclops system's requirement when accessing its library files, it was based upon the programme described in section (A3.2.2.1). With over twelve thousand lines of code

in 24 source and header files, the programme represents the culmination of all the techniques that were tried during the research and that were found to demonstrate at least partial success in terms of an improvement with some aspect of the optophonic process.

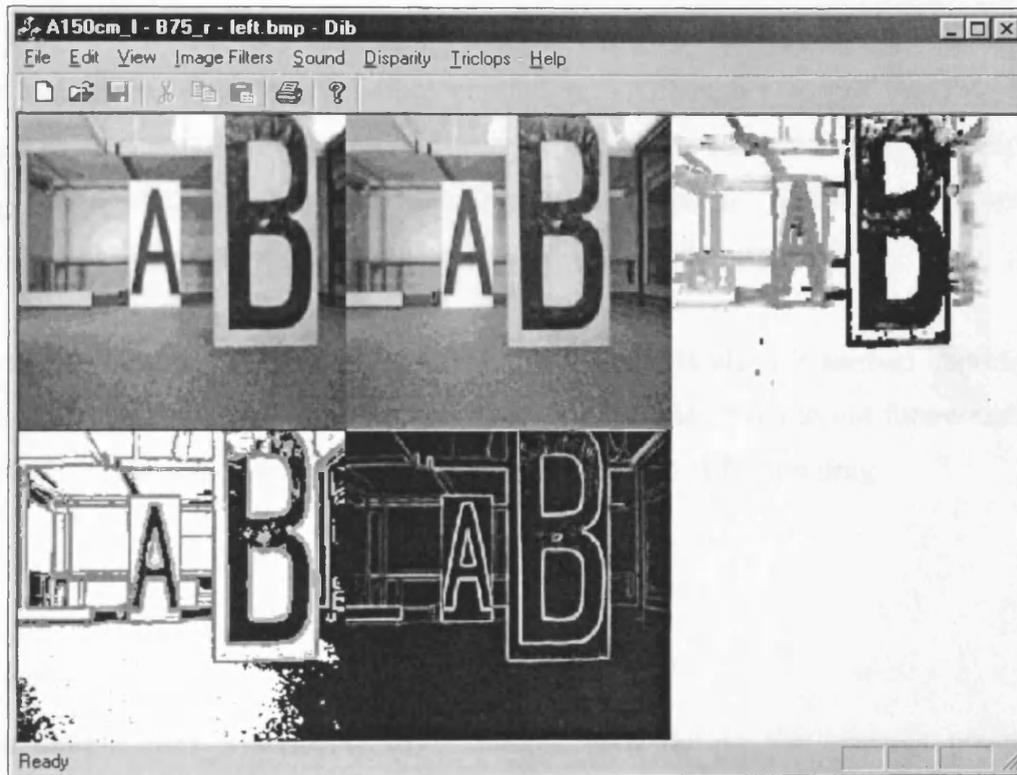


Figure (A3.5b) – Screenshot of the real-time stereo optophone programme portraying the capturing of images and the generation of cartoon depth maps. Other images shown, and used during stereo processing, include a cartoon image (bottom left) of the left frame (top left) and an edge image (bottom middle) created from the right frame (top middle).

The option controls for this programme were designed in a similar fashion to that of the stereo algorithm programme, written in Borland Builder C++ and described in section (A3.2.2.1), and were accessible through dialog windows similar to those shown in figures (A3.3a)-(A3.3c). The only difference being the inclusion of a *Triclops dialog window* that provided control over the various image capturing

features, such as the image sizes and the application of Triclops pre-processing filters, as well as the option to capture video to disk for later playback.

A3.3. Hardware & software problems

As expected, during the course of the research numerous difficulties were encountered that needed to be remedied before continuing. Although many of these problems were, at most, simply annoying and were on the whole easily solvable, they did take time and so detracted from the main aim of the research. Namely, that of making progress in the quest for improvements to the optophonic mapping!

Considering the time spent solving some of these difficulties it seemed important to describe a few of the slightly more problematic events, if not to aid future research, then to at least demonstrate that the work was not always plain sailing.

A3.3.1. Camera misalignment

The three-camera system, called Triclops, used during the research played an important role, and was invaluable for testing modifications to stereo algorithms. Originally stereo testing began with two-camera stereo techniques, later progressing to three-camera systems (mainly using the third camera for verification of candidate matches). At this time a disturbing problem presented itself.

Using the third camera, rather than providing enhanced depth maps, appeared to generate more errors than the equivalent two-camera system, as shown in figure (A3.6a). There could only be two possible explanations. Either the computer program had some logical errors or one of the cameras was misaligned.

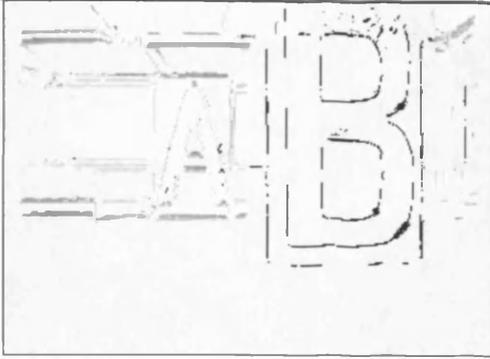


Figure (A3.6a) – Three-camera edge depth map showing a large number of errors due to the misalignment of the top camera.

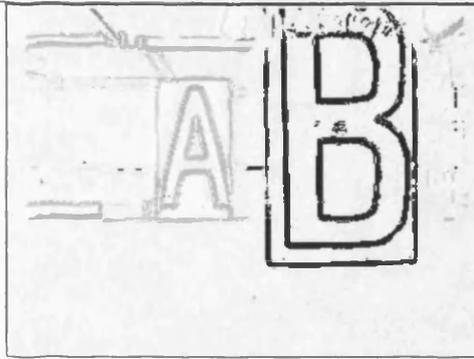


Figure (A3.6b) – Three-camera edge depth map after the necessary corrections were made for the misaligned camera.

As indicated by the title of this section, it was the latter of the two – a camera misalignment. To determine what sort of misalignment exists is quite a simple task, requiring a quick comparison (by hand) between the three stereo images of a selection of points, each with differing depth. Unfortunately, the task of solving a camera misalignment is not always as simple. For example, if one of the cameras had been rotated then it would point in a different direction to the others, causing objects at different depths to be misplaced by differing amounts.

Figures (A3.7a)-(A3.7c) – The black crosses marked in these figures, bottom-left of the 'A' & 'B' signs, were used to check the alignment of the three cameras in the Triclops camera system. These image points could effectively be used since the distances of the 'A' & 'B' signs were known to be at a distance of 1.50m and 0.75m, respectively, from the observing cameras.

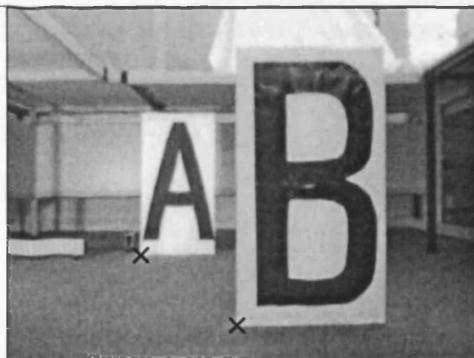


Figure (A3.7a) – Misaligned top image of the set of three that comprise the Triclops system.

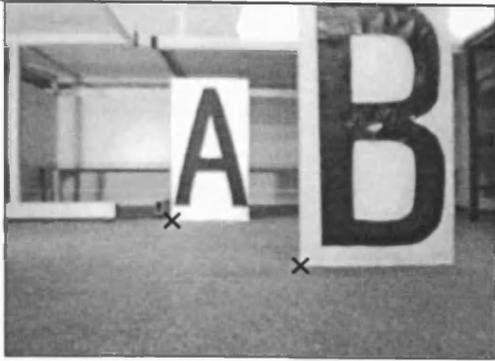


Figure (A3.7b) – Image obtained from the left Triclops camera.

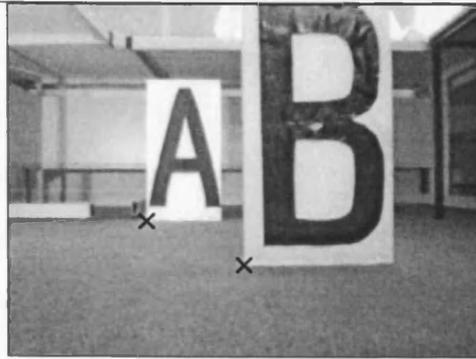


Figure (A3.7c) – Image obtained from the right Triclops camera.

Normally when calibrating stereo cameras it is necessary to use specially prepared grids that are placed at a specific distance from the cameras and used to calculate any misalignments. However, in the case of a three camera system, where it can be assumed that two of the cameras are correctly calibrated (due to the accuracy of the generated two camera depth maps) the task can sometimes be a little easier.

As an example of finding camera misalignments, figures (A3.7a)-(A3.7c) represent a scene consisting of two boards or signs with the letters 'A' & 'B'. The 'A' is at a distance of 1.50 metres from the cameras, and the 'B' is at 0.75 metres. For perfect camera alignment the disparity between the three images of any point on the 'A' should be exactly half that of the 'B'. In these three figures two points have been marked, by black crosses, that lie on the bottom-left corners of the 'A' & 'B'. Table (A3.1) and figures (A3.8a) & (A3.8b) show the coordinates, disparities and misalignments found between these sets of points.

Point Location	Top Camera	Right Camera	Left Camera	Correct Disparity (Left-Right Cameras)
Bottom-Left 'A'	(89, 168)	(92, 147)	(110, 147)	18 pixels
Bottom-Left 'B'	(154, 216)	(157, 177)	(193, 177)	32 pixels

Table (A3.1) – Coordinates of the crosses marked in figures (A3.7a)-(A3.7c) used to calculate camera misalignment.

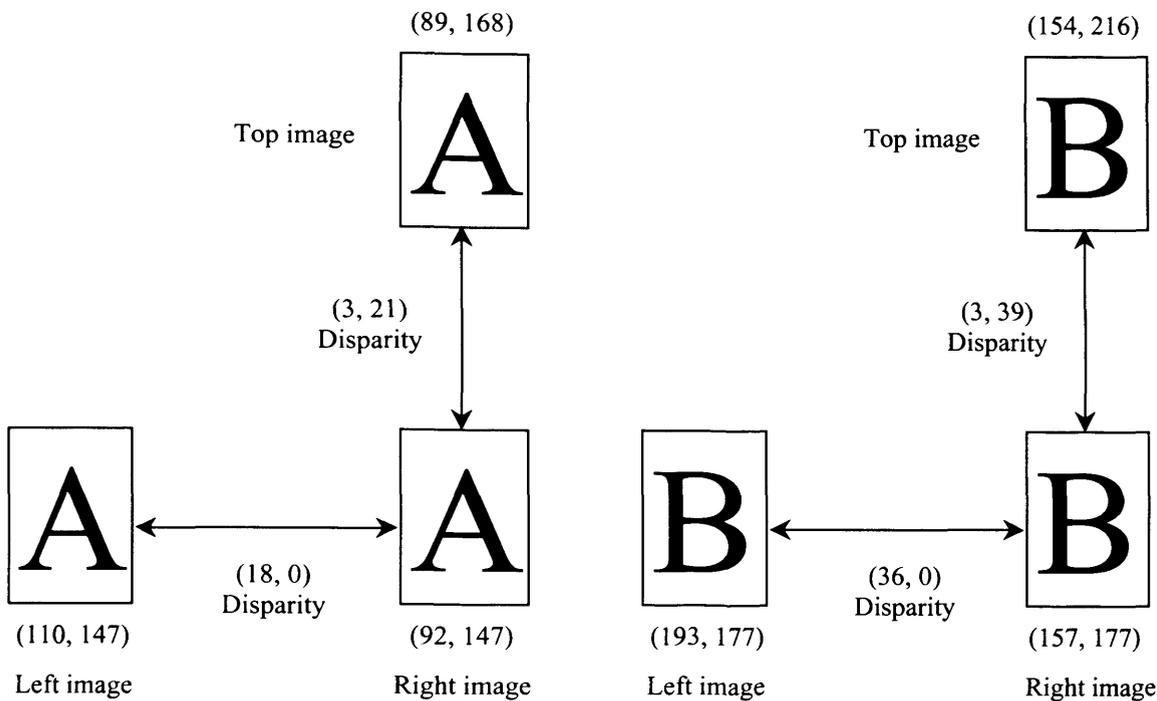


Figure (A3.8a)

Figure (A3.8b)

Figures (A3.8a) & (A3.8b) – These figures correspond to the coordinates of the crosses seen in figures (A3.7a)-(A3.7c). The disparities are also shown in the above diagrams.

From table (A3.1) and figures (A3.8a) & (A3.8b) it can be seen that the left and right stereo cameras are inline with each other. On the other hand, the top camera, for all depths, appears to be 3 pixels out of alignment, in both the vertical and horizontal directions. This means that the top camera had only translated, rather than rotated. Consequently, a simple solution is to capture the images and then adjust all calculations by three pixels in both the x and y directions. After making these adjustments to the stereo computer algorithm all depth maps produced are of a much higher standard, as shown in figure (A3.6b).

A3.3.2. Programme development

Although it was not so much of a problem in terms of something not working properly, the task of developing the programmes, or rather, the learning involved prior

to and during the creation of the programmes was quite severe and took considerable time.

When starting the research the programmes used were written in 'C' on a basic 486 PC. However, this neither had the necessary cards to support capturing from camera nor the speed to accomplish what was deemed necessary for the investigation into real-time conversions of images-to-sound. It was at this point that the work moved onto the Unix based Silicon Graphics O2, which was by far superior in terms of speed and that it was designed for the processing and capturing of images. However, this meant learning to use Unix and to programme in Gnu C++ using the *X Windows System* (and to some extent OpenGL). After extensive use of the Silicon Graphics website libraries and various Silicon Graphics programming guides a fully working real-time image-to-sound system was produced. This provided an extremely useful base on which to test numerous algorithms and ideas.

At this stage various image processing algorithms began to show promise, and after a great deal of optimising with respect to the code, the programmes were ported back onto a Pentium PC to provide a slightly smaller and more mobile system. Any new algorithms were first tested on the Silicon Graphics, and if found to show promise they would then be optimised (in terms of speed) for the PC.

Programmes had been written in various straightforward C++ compilers, however it was decided that a graphical user interface should be incorporated, and that Borland Builder C++ would provide a suitable frame on which to begin. Since each compiler uses slightly different commands and that Builder also employs a graphical designer for building applications, this required an additional learning period. Furthermore, until then all programmes written on the PC worked through DOS and accessed the sound and video cards directly, however, the Borland Builder programmes were being created within Windows 95. Since Windows uses a system known as messaging to access hardware such as a soundcard, this required further time to study.

Another similar transition was required after purchasing the Triclops camera system since this required the use of Microsoft Visual C++, which although laden down in helpful features, used a completely different layout and command system.

During this necessary, but lengthy, period of learning much progress was made with respect to the optophonic mapping, however it must be said that each session did require time that could have otherwise been spent researching the optophonic process.

A3.3.3. Optimisation

Although in many cases the solution was found to be fairly obvious after the event there were a number of occasions where speed was an important issue. Many of the programmes written were required to process multiple stereo frames per second to generate a depth map, not forgetting the additional time required to perform the conversion from image to sound. For instance, in the case of the sound processing a vast increase in speed could easily be made by building a large array as a look-up table that contained all possible sine values to be used. In this case, whilst generating the sound, rather than continually performing complex calculations the value could simply be obtained by accessing the appropriate cell in the look-up array.

One particular occasion where a considerable speed increase was required occurred whilst creating an adaptive threshold for the process known as 'cartooning', described in chapter (4.3.5). The original method for setting a threshold merely chose a pixel intensity beyond which all values would be set to white, otherwise black. This was found to be unacceptable since not all scenes have similar intensities. Selecting a percentage threshold and applying it to a histogram of the intensity values for the captured scene can provide a more suitable and adaptive alternative. However, this requires scanning the whole image twice – once for the histogram, and again to create the actual cartoon image. Consequently, the overall frame rate is greatly reduced.

A quick and easy solution for increasing the speed of this process is to skip pixels while creating the histogram. This provides an approximate histogram, which is more than adequate for its purpose, at a fraction of the time. For instance, if the original image is 320x240-pixels in size, then by skipping three pixels in four both horizontally and vertically the total number of pixels scanned is 80x60 (a sixteenth of the original size).

In the case described the process would be almost 16 times faster than scanning the whole image to generate the histogram. Using this technique there is almost no perceptual difference in frame rate between the cartooning technique with or without the adaptive thresholding, however the visual improvement when using the thresholding is noticeable when capturing directly from a moving camera.

Appendix 4. Publications

During the course of the research four papers were written and submitted. Three of which have been published and the remaining paper recently submitted. These papers are listed in the 'Reference' chapter, and are as follows:

[CapPic00a] – Capp, M. and Picton, P., (Feb.-Mar. 2000). 'Fast, Low Resolution Edge Depth Maps and their Application to a Blind Mobility Aid', *International Conference on Computer Vision, Pattern Recognition and Image Processing*, Atlantic City, USA, **CVPRIP-12**, pp. 248-251.

[CapPic00b] – Capp, M., & Picton, P., (Jun. 2000). "The Optophone: an Electronic Blind Aid." *Engineering Science and Education Journal*, vol. **9**, no. 3, pp. 137-143.

[CapPic00c] – Capp, M. & Picton, P., (Aug. 2000). 'An investigation into stereo vision as a modification to optophonic mappings from scene-to-sound.' *10th International Mobility Conference*, Warwick, 4th-7th Aug. 2000.

[CapPic00d] – Capp, M. & Picton, P., 'Relaying Scene Information to the Blind via Sound using Cartoon Depth Maps.' Submitted to *IEE Proceedings – Vision, Image and Signal Processing*, 2000.

These papers are included in the following pages.

Fast, Low Resolution Edge Depth Maps and their Application to a Blind Mobility Aid

Michael Capp and Prof. Phil Picton
University College Northampton, UK

Abstract

In the quest for a suitable transformation from images to sound to enable blind people to navigate and read, a technique has been developed that produces a low resolution, but extremely fast, stereo edge depth map. In this paper the logic behind the research is described, and a description of this stereo technique given with examples. The technique produces a depth map whereby the distance to an object is represented by the thickness of the lines in the image.

Introduction

Blind mobility aids [BenBen63, Kay84, Mei92, Mei99, NieMahMea87, Ohe94, ShoBorKor98, WarStr85] such as the Optophone [Fou24], invented by Fournier d'Albe for converting light intensities or images into sound, are often plagued by the problems of information clutter or overload. Although blind mobility aids have progressed considerably since the time of Fournier d'Albe's first Optophone, this problem still remains. The reasons for this are apparent when the bandwidth reduction necessary is considered when converting real-time image flow (vision) into forms acceptable for other sensory modalities, such as our haptic or auditory senses. As a result of this, many electronic blind aids are both mentally tiring and stressful to use.

An important first question in designing a blind mobility aid is therefore what information can be discarded from a scene, whilst keeping the essentials to allow a blind user to comfortably navigate their surroundings? It is also important to bear in mind that the human brain is far superior in filtering image/sound data than any present day computer, so there is no need to throw away all of the redundant information.

Since the invention of the first electronic blind aids, almost one hundred years ago, there have been many suggestions on how to reduce the complexity of an image. Some of the simplest methods consist of the reduction of input image size and the number of colours (or rather grey levels), image segmentation [Ohe94], object detection and recognition, and edge detection. Alternatively, a different approach to this problem is through the use of ultrasonic beams [BenBen63, WarStr85, ShoBorKor98] to give the user an

indication of the objects and obstacles in their immediate vicinity (normally a range of approximately 3 to 4 metres in a wide arc in front of the user).

However, in this paper a slightly different approach to image information reduction is proposed, namely, stereo depth maps.

A computationally inexpensive edge depth map

By using multiple cameras or a sequence of images from a moving camera, a stereo depth map can be constructed [Ber97, Fua93, MarDurCha85]. In so doing the quantity of information conveyed to the user is reduced by representing distant objects with a lower amplitude (or zero) sound. This helps in a mobility aid where generally blind people do not need to 'see' what is at the end of a street, but are more concerned with objects and obstacles in their immediate vicinity.

There are a great many stereo techniques that have been tried, each of which with their own pros and cons. However, the technique proposed, which is founded upon a time-varying edge detector [Pic89], generates a very fast, albeit low resolution, edge depth map. This method is based in part, on a stereo-optical edge detector that was proposed by Avery Johnson [Joh63] in 1963, who believed it would be possible to use a small mosaic of photosensitive elements to locate nearby objects (contrast boundaries). The proposed system, while moving with a translational motion, would detect only the closest objects in the scene. As the device containing the array of photosensitive cells moved, nearby objects would be observed as moving with a greater velocity than distant objects. Consequently, the light from the closest objects would cross the greatest number of the device's array elements, whereas distant objects would not cross any cells in the mosaic (in the short term), hence not registering.

Assume that a pair of stereo images, Frame1 and Frame2, have a size in the region of 320x240 pixels or less. A low pass filter is run across the two images (alternatively, if the images are large, they can be reduced by taking the mean values), followed by a large Prewitt or Sobel edge operator with one added constraint. All edge pixels above a minimum threshold value are set to 1; the others are set to zero. Finally, the depth map is generated by setting the current pixel in the output image to 1 if the equivalent pixel in Frame2 is 1, and Frame1 is zero. Otherwise, the output is set to zero.

This can be summarised as:

```
If (S(frame1) > min_threshold)
    Then S(frame1) = 1
Else S(frame1) = 0
If (S(frame2) > min_threshold)
    Then S(frame2) = 1
Else S(frame2) = 0
If (S(frame2) == 1 and S(frame1) == 0)
    Then Result = 1
Else Result = 0
```

Where S represents a 3x3 (or preferably larger) Sobel or Prewitt edge operator, or any other suitable edge detector that is capable of producing relatively 'thick' edges. The result is that depth is proportional to the thickness of the edges in the resulting image, where a one signifies an edge.

The idea behind this technique is that a distant object will generate a line that lies in the same position (or nearly the same position) in both frames. Therefore, if the two images were to be overlaid, a large or complete overlap of the two lines would be seen. Hence, the procedure described earlier would either reveal a set of zeroes, or a very thin line along the region where the two edges did not overlap. This effect would reveal wider edges for objects nearer to the camera(s), since the greater the disparity between the edges, the less they overlap.

Limitations and Possible Remedies

Of course, there are a number of restrictions and considerations that must be taken into account when using this algorithm, but for use as a mobility aid, these are minor in comparison to the speed of computation. Nonetheless, these limitations still exist.

The first and probably the most obvious drawback with this technique is the loss of edges in the depth map that lie parallel to the epipolar line. However, this is a common failing of many stereo depth maps, and is similar in nature to the aperture problem of cameras. For example, consider a simple scene with a long horizontal line that more than fills the view of an observing camera. Without texture in the scene it would be hard to identify any horizontal motion of the line. In the case of the technique presented here, if it encounters an edge that lies along the epipolar line, then it cannot determine the line's disparity, so the line will not be shown in the final image. See

figure 4 for an example.

Another problem that reveals itself in this and other stereo systems arises from repetitive patterns in the scenery. In this case, if two or more parallel lines are encountered in close proximity, then this technique will usually remove a line, or part of a line (depending on the disparity at that location), whilst generally displaying the rest. This will only happen when the disparity is great enough to cause adjacent lines in the stereo image pair to overlap.

Not so much of a limitation, but as a consideration, is the maximum recognisable disparity that this procedure exhibits. The maximum disparity that this technique registers correctly, in terms of pixels, is equal to the width of the edges in the image after applying the edge operator. (Note – any disparities greater than the width of the edges, i.e. from objects that lie very close to the cameras, will be seen to have the same disparity as that indicated by the line widths themselves). As a direct consequence of this, the stereo cameras are restricted to being positioned very close together so that the maximum stereo disparity that is encountered is generally not too high. Or, in the case of a single camera for a monocular system, scenes would have to be sampled at a very high frame rate, taking consecutive frames as the stereo pair of images. This limits the maximum possible movement of any object in the scene between frames. Fortunately, with the speed of operation of this algorithm, computation at a high frame rate is quite acceptable.

However, the use of a low pass filter before applying the edge detector to the images can lessen the effect of the above restriction. This is achieved by a widening of the edges in the images, as well as aiding in the removal of excess texture. Alternatively, reducing the size of the initial images, say from 320x240 to 160x120 pixels, by taking the mean value of every 2x2 block can help in two ways. Firstly, by performing a very effective low pass filter, and secondly, by further reducing the necessary computation time for the remainder of the procedure. In addition, the resulting depth map is more pleasing to look at due to the ratio of the maximum line width to image size. As the image is reduced in size, the thickest lines (those depicting the greatest disparities, assuming the lines retain a constant width) are represented more clearly. For example, a line of 10 pixels in width would be more noticeable in an image of size 160x120 pixels than in one of 320x240 pixels. This assumes that the image lines will demonstrate a near constant thickness before

and after the process of image reduction, which is near to the truth, since the process of mean value reduction effectively spreads the majority of edges.

The range of disparities that this procedure accurately displays can be further increased by using a method of edge thickening before proceeding with the generation of the depth map.

An additional limitation that only arises when using a monocular camera system is the need to restrict camera motion to purely translational movement. Again, this is a common failing of monocular camera systems, which can often result in a reversed depth map, or worse. For example, if the camera undergoes a rotation about the horizontal (such as the holder of the camera turning round), then distant objects will appear to have a greater disparity than those closer at hand.

Results

The maximum line width that was generated by an enlarged (5x5) Prewitt operator in the example images was six pixels. This means that the depth map in figure 4 has six (seven if the lack of a line is included – white) possible ranges or depths that correspond to the six possible line widths. However, as stated above, it is possible to include a line thickening stage in the procedure that could improve this resolution. Similarly, by further increasing the size of the initial low pass filters and edge operators, at the cost of a greater overall processing time, it is possible to increase the maximum line width, and hence, improve the depth map resolution.

The depth map that results from this procedure is quite adequate for the purpose of developing a mobility aid. However, most stereo edge depth maps consist of lines of varying intensity, whereby the intensity is directly proportional to the distance to the edge of a real world object, instead of varying edge thickness. Consequently, it was felt necessary to include some sort of comparison.

To make this comparison the line width scale has been converted into an intensity-based scale (figure 4). Since there are seven different line widths (0, 1, 2, 3, 4, 5 & 6 pixels) in the depth maps, the corresponding pixel intensities range from 0 to 255: 0 (foreground), 42, 85, 127 (middle distance), 170, 212, 255 (background). Zero corresponds to a black line 6 pixels in diameter, and an intensity of 255 represents a white pixel, or the lack of a line.

As a guide, the depth map obtained from a simple stereo edge detector has been included (figure 3).

This detector uses the Sobel edge operator to obtain the local edge magnitudes and orientations from the stereo image pair. Once found, matches are made by searching for the candidates that have near identical edge properties within a given disparity range. A second constraint that further reduces the number of incorrect matches is through the use of previous disparities. It is often the case that several edge pixels are located from the same real world object. Hence, if two likely candidates are found, then priority is given to the one that most closely matches the previous disparity that was correctly located.

A candidate match for this second stereo procedure is one in which the difference between edge orientation is less than 30 degrees, and the magnitudes do not vary by more than 30 (from an intensity range of 0-255).

Image to Sound

Once the depth map has been successfully created, it is further reduced in resolution to 64x64 pixels, which can readily be converted into sound, column by column. This rapidly generated sound, which lasts for a period of about one-second, makes use of stereo effects and frequency changes to give the impression of object location. Similarly, changes in amplitude are used to represent the relative distance to an object in the real world. Hence, even a very low-resolution depth map adequately provides the user with enough information to locate objects in their immediate vicinity, as well as allowing them to read large text.

Conclusions

In this paper a very simple technique for generating basic edge depth maps has been proposed which, although limited in its resolution, is capable of producing edge depth maps in real-time on a standard PC. As a result, it is ideal for navigation aids where very fast but simple depth maps are required. Work is currently being undertaken on alternative methods which produce a higher quality depth map but inevitably are slower. This trade-off between speed and quality will ultimately determine the type of navigation aid that can be produced.

References

[BenBen63] – Benham, T. A., & Benjamin, J. M., Jr., (1963). "Active energy radiating systems: An electronic travel aid." *Proc. of the international congress on technology & blindness*, The Amer. Found. for the Blind: New York, 1, pp. 167-176.

[Ber97] – Bergendahl, J. R., (1997). “A computationally efficient stereo vision algorithm for adaptive cruise control.” MSc thesis, Dep. of Elec. Eng. & Comp. Science, Massachusetts Institute of Technology.

[Fou24] – Fournier d’Albe, E. E., (1924). “The Moon Element – An introduction to the wonders of Selenium.” T. Fisher Unwin Ltd. London.

[Fua93] – Fua, P., (1993). “A parallel stereo algorithm that produces dense depth maps and preserves image features.” *Machine Vision and Applications*, vol. 6, pp. 35-49.

[Joh63] – Johnson, A. R., (1963). “Passive Systems: A proposed stereo optical edge detector.” *Proc. of the International Congress on Techn. & Blindness*, The Amer. Foundation for the Blind: New York, vol. 1, pp. 183-186.

[Kay84] – Kay, L., (1984). “Electronic aids for blind persons: an interdisciplinary subject.” *IEE Proceedings*, vol. 131, pt. A, no. 7, pp. 559-576.

[MarDurCha85] – Marshall, S., Durrani, T. S., & Chapman, R., (March 1985). “A binocular stereo matching algorithm.” *Computer & Control Colloquium on ‘Vision Systems in Robotic and Industrial Control’*, The Institution of Electrical Engineers, Savoy Place, London.

[Mei92] – Meijer, P. B. L. (Feb. 1992). “An experimental system for auditory image representations”, *IEEE Trans Biomed Eng.*, vol. BME-39, no. 2, pp. 112-121.

[Mei99] – Meijer, P B L. (Accessed April 1999). http://ourworld.compuserve.com/homepages/Peter_Meijer. [Online].

[NieMahMea87] – Nielson, L., Mahowald, M., & Mead, C., (1987). “SeeHear.” *Proc.: Image Analysis*, 5th Conf., Stockholm, 1, pp. 383-396.

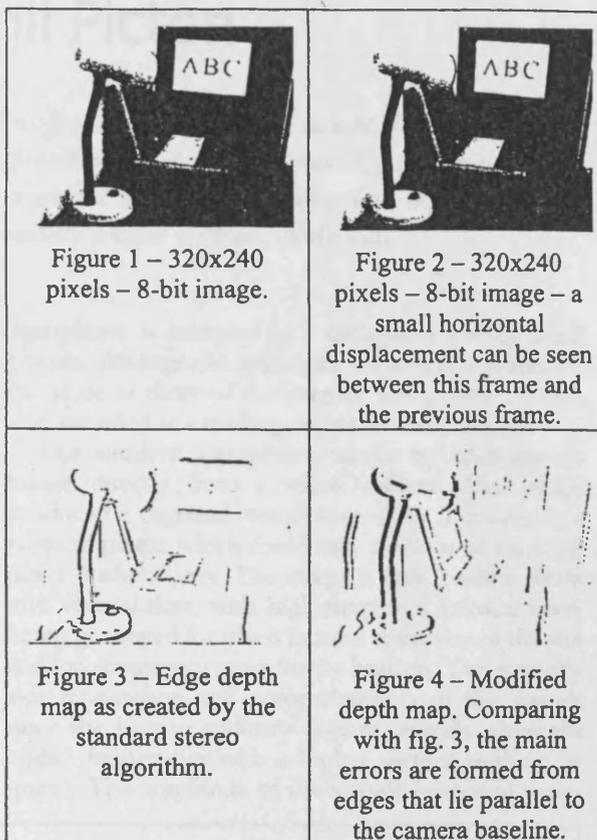
[Ohe94] – O’Hea, A. R. (Apr. 1994). “Optophone Design: Optical-to-Auditory Vision Substitution for the Blind.” PhD Thesis, Open University, UK.

[Pic89] – Picton, P. D., (Jul 1989). “Tracking and segmentation of moving objects in a scene.” *IEE 3rd Internatl. Conf. on Image Processing and its Applications*, Conf. Pub. No. 307, pp. 389-393.

[ShoBorKor98] – Shoval, S., Borenstein, J., & Koren, Y. (Nov. 1998). “The Navbelt – A Computerized Travel Aid for the Blind Based on Mobile Robotics Technology.” *IEEE Trans. on Biomed. Eng.*, vol. 45, no. 11, pp. 1376-1386.

[WarStr85] – Warren, D. H., & Strelow, E. R., (Editors) (1985). “Electronic spatial sensing for the blind.” ISBN: 90-247-3238-7, 521pp.

Appendix: Examples



The optophone: an electronic blind aid

by Michael Capp and Phil Picton

This article presents a short but detailed description of the optophone—its origins as a reading device for the blind, the various stages of its development, and the possibility of its use as a mobility aid for the blind. Recent research into the use of stereo vision is described as an aid to information reduction, in the hope of remedying the problems of information overload that commonly plague electronic blind aids.

The principles of the optophone have been around for nearly 100 years, the first working device having been built as long ago as 1912. The obvious questions are therefore why only a handful of people have heard of it and why it is not used. A three-year project at University College Northampton (UCN) has been sponsored by the Engineering and Physical Sciences Research Council (EPSRC) to try to answer these questions and to see if the idea can be improved with the aid of new technology and some help from volunteers. The first step has been to look back at the various optophones and similar devices that have been tried in the past. Included in this list of devices is the modern optophone^{1,2} that was invented by Dr. Peter Meijer and patented by Philips in the Netherlands. Meijer's

optophone is intended as a navigation aid for blind people, although the principles on which it works are the same as those of the original optophone³, which was intended as a reading device for blind people.

The modern optophone works by scanning an image directly from a video camera. The image produced is digitised, which means that it consists of a series of pixels, which could have a colour or could be just a shade of grey. The image is then broken down into vertical slots, with high-frequency musical notes being generated for pixels located at the top of the slot and low-frequency notes for the bottom. This instantly aids recognition and comprehension of the sounds since the human auditory system naturally associates higher frequencies with a higher vertical position in space¹. The amplitude of the sounds generated varies

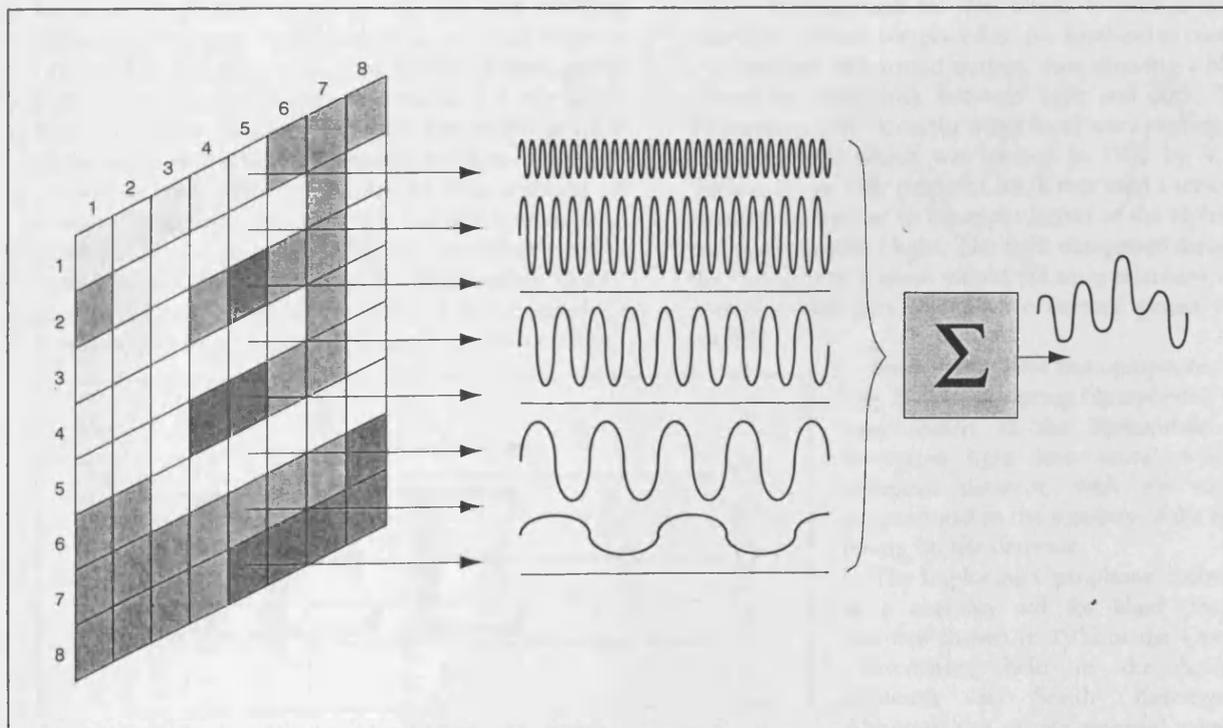


Fig. 1 Simplified diagram of Meijer's optophone system



Fig. 2 Blind man using the 'Exploring Optophone'. Fournier d'Albe can be seen standing on the left of the picture. (From Ref. 3)

proportionally to the intensity of the light captured by the video camera.

Fig. 1 demonstrates a simplified version of a transformation from image to sound. This is called a *slot mapping* or the *piano transform*, aptly named since it can be represented as a vertical piano keyboard scrolling across a 2-D image and generating musical notes as output. The slot is moved across the image from left to right. Now the sound output generated at any given time is a musical chord solely dependent on the portion of the scene that is visible (through the slot).

One of the main problems of this method of converting pictures into sounds is that the human ear is nowhere near as good as the human eye for processing information. This means that the information from a picture has to be dramatically reduced. For example, if it is assumed that the video image is a 24-bit colour

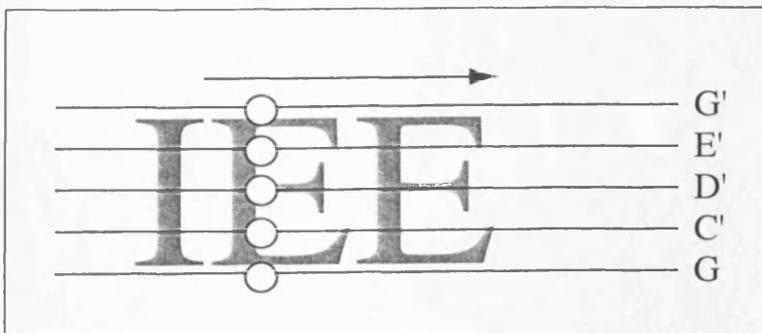


Fig. 3 Notes produced by Fournier d'Albe's optophone

VGA image, displayed at 25 frames per second, the number of bits per second is:

$$640 \times 480 \times 24 \times 25 = 184\,320\,000 \text{ bits per second}$$

The bandwidth of audible sounds is about 15 kHz, allowing a bit rate of 30 000 bits per second, so the information has to be reduced by a factor of 6144! One way this can be achieved is to reduce the size of the image and to reduce the number of bits per pixel. For example, if a monochrome image of 64×64 pixels is used with 4 bits per pixel and 1 frame per second, the bit rate is reduced to 16 384 bits per second, which is within the audible range. In fact these are the figures used by Meijer in his prototype implementation of an optophone, which will be described later.

A brief history

The word 'optophone' was first used by Dr. E. E. Fournier d'Albe³, who in 1910 was appointed Assistant Lecturer in Physics at the University of Birmingham in England. It was here that he set up a laboratory to look into the properties of selenium. Selenium had been discovered by Berzelius in 1817 and in the years that followed it became apparent that selenium

was photosensitive, reacting to light in such a way as to vary its conductivity. Possibly unknown to Fournier d'Albe, a few devices had already been invented, such as the Elektroftalm⁴ and the Photophonic book⁵. The Elektroftalm, originally created in 1897 by Noiszewski, was a mobility aid for the blind. It used a single selenium cell that was placed on the forehead to control the intensity of a sound output, thus allowing a blind person to distinguish between light and dark. The Photophonic book on the other hand was a reading aid for the blind, which was created in 1902 by V. de Turine. A specially prepared book that used a series of transparent squares to represent letters of the alphabet was passed under a light. The light that passed through the transparent squares would fall on a selenium cell, which would vary the electric current passed to a speaker.

Fournier d'Albe's first optophone (see Fig. 2), the 'Exploring Optophone', was very similar to the Elektroftalm. It converted light into sound with a selenium detector, with the sound proportional to the intensity of the light falling on the detector.

The Exploring Optophone, intended as a mobility aid for blind people, was first shown in 1912 at the Optical Convention, held in the Science Museum at South Kensington. Although this device received a lot of

good press coverage, it was criticised by Sir Washington Ranger, a blind solicitor, who said: 'The blind problem is not to find lights or windows, but how to earn your living.' As a result of this criticism Fournier d'Albe turned away from mobility aids, and in 1913 invented the Reading Optophone. It would scan across a line of text, with guidance from the user, and produce a higher intensity sound for the lighter areas of the text, producing no sound for the black characters. The optophone used a vertical row of five spots of light. When the light passed over the page, the reflection varied according to the character, or lack of character, scanned. The reflected light was detected by a row of selenium cells, which modulated the sound output, each spot of light producing a different musical note. The idea for this was that a musical 'motif' would be produced from the reading of text, which would be far more pleasing to the ear than a set of unfamiliar sounds.

There was an inevitable gap in the research during the First World War. However, in 1918 the firm of Barr & Stroud took over and developed the 'black sounding' optophone. Due to the lower ratio of black print to blank paper on a page, this created fewer sounds and was considered to be easier to learn. One expert student of the device, Miss Mary Jameson (see Fig. 4), attained a reading speed of 60 words a minute, although the average was generally a lot lower. At this stage the optophone seemed on the verge of success. It was being manufactured in significant numbers and had a great deal of publicity, even appearing on the cover of *Scientific American* in 1920. However, by 1924 it had all but disappeared. The main reason for its disappearance seems to be that it was not an easy device to use. It required training, it was quite stressful to use, and reading speeds were quite slow.

Over the years, many variations of the reading device have been tried. Just after the Second World War, the Veteran's Association in America commissioned a project to look into reading devices for the blind. This project was set up at the Haskins Laboratory⁵ and ran for 30 years, during which time a number of variations

on the optophone were tried by F. S. Cooper and his colleagues. Eventually the project came to the conclusion that the ideal reading aid for the blind is an optical character reader (OCR) linked to a speech synthesiser. Of course, by that time (1974) such devices were starting to appear. This was the final blow to the optophone as a reading aid for the blind.

Although the research into reading aids for the blind has more or less ceased, the search for mobility aids or obstacle detectors continues. Since the 1940s, a number of electronic mobility aids have been proposed. One of the more bizarre devices was the Radioactive Guider⁶. This device produced a 'beam of radiations' by using a small particle of radioactive material. The radiation reflecting off a nearby object could be detected via a Geiger counter and used as an indication of the distance to that object. How far they managed to get with these trials can only be speculated!

Currently one of the best-known mobility aids is the Sonic Guide⁷ of Leslie Kay, in which echolocation employing ultrasonic waves is used to present the user with an audio representation of their surroundings. With this device users can learn to replace partially their lost sense of sight as they have been shown to have an improved spatial understanding and awareness.

The question that arises is whether the principle of converting light to sound as in the reading optophone can be extended to navigation. A number of people have thought of this and even got as far as patenting devices. These are essentially variations of the original travelling optophone or reading optophone, the main differences being in the scene-to-sound mapping and the output qualities, such as resolution. Known attempts at optophonic mappings include the following:

- Fish⁸ (1976) used frequency to map vertical position, and binaural loudness difference for horizontal position. The sound heard at any instant depended on the brightness gradient of one point in the scene and the scene was scanned by the single point in



Fig. 4 Miss Mary Jameson at the Reading Optophone (From Ref. 3)

raster fashion. This form of mapping is classified as a *point mapping*.

- Dallas⁹ (1980) mapped vertical position in the scene to frequency, horizontal position to time, and brightness to loudness. The mapping used is an example of the piano transform, which was re-discovered independently by O'Hea¹⁰ and Meijer¹.
- Kurcz¹¹ (1981) used another point mapping, in the form of a hand-held device called a heliotrope, which sensed the light output from only one point in the scene, that point being controlled by the user. Again, the sound output was a function of the light intensity.

The two most significant pieces of recent work are by Adrian O'Hea^{10,12} and Peter Meijer^{1,2}. O'Hea spent 8 years working on his PhD on optophones at the Open University in the UK. In his thesis he tried to extract the best methods by comparing other attempts, and this is therefore the first major study of optophones. One major disadvantage at that time was an inability to produce real-time sounds, and so his thesis was largely untested. Unfortunately he was unable to take this work further due to his untimely death shortly after completing his thesis.

One of O'Hea's suggested mappings was the same piano mapping as used by Dallas, although O'Hea was unaware of this at the time. He found this mapping unsatisfactory. For example, a wide light shape on a dark background produces a mapping which is equivalent to trying to convey two notes on the piano (the edges of the shape) by playing all the notes in between. A second slot mapping that he tried mapped edge orientation to musical (circular) pitch, and position along the slot to interaural intensity difference. He continued by suggesting a cosine, polar piano and even a free-field patch transform in which the scene is split into segments of interest that are dealt with individually.

O'Hea proposed a method, called the 'theoretical performance test' (TPT), for evaluating the mappings and gave examples. This was necessary because, as mentioned above, it was impossible to generate real-time sounds. The TPT is made up of the following six stages that make it possible to find flaws in prospective mappings:

- (1) Obtain a digital image.
- (2) Using the trial mapping convert the image to a sound file.
- (3) Using a model of the known inaccuracies of human hearing, modify the sound so that it represents the way in which a human would perceive it.
- (4) Recalculate the scene using an inverse of the mapping and the modified sound.
- (5) Criticise the recalculated scene visually, by comparison with the original.
- (6) See if there emerge any clues to a better mapping.

Although there were a number of limitations with the

TPT method, it did enable O'Hea to make a number of extensive evaluations and comparisons between prospective mappings that would otherwise have been impossible without further backing and equipment.

At the same time, Peter Meijer was developing an optophone at the Philips laboratory in Eindhoven. He used the piano transform in a similar way to that of Dallas, O'Hea and even Fournier d'Albe. Not only has Meijer created a working real-time device, which he has called 'the voICe', but he has also produced an excellent web site¹³ that describes his work and provides links to relevant sites. To date this work represents the furthest that anyone has gone in optophonics, but its effectiveness as a navigation aid has still to be proved since extensive trials have yet to be carried out.

Current work

Adrian O'Hea very firmly believed that the optophone was the way forward. He argued that an optical input supplies the device with the information available to a sighted person. Echolocation with ultrasonic waves, for instance, cannot see anything far, or through glass, or on paper. Also, hearing seems to be able to convey a great variety of information and headphones seem the least invasive form of output device. He also noted that a device with an optical input should not be considered merely as a mobility aid. If a device can report shapes, it can be used for reading, for example, signs and notices in the street. Our research project at UCN stays true to O'Hea's beliefs but realises the importance of depth information, which is readily available to a sighted person with two eyes. Therefore, unlike Meijer's optophone, stereo image processing has been used as part of the system.

Sound outputs

Various techniques for presenting the output from the optophone in a more pleasing manner have been investigated. Also various methods have been tried to prevent information overload, whilst retaining enough scene information for the user to navigate comfortably and easily (ideally) around their local environment (not forgetting that the human brain is far superior in filtering image/sound data than any present-day computer).

The basic display method used by an optophone is a monaural sound output, with the horizontal displacement (across the scene) represented as a function of time. In Meijer's optophone a click is used to indicate the start of each frame. The vertical dimension of the scene is represented by frequency of sine waves, since the human auditory system naturally represents (to a certain degree) height with higher frequencies. For example, the human ear recognises (roughly) a logarithmic increase in frequency as a linear increase in height of the sound source^{1,8}.

An immediate improvement has been made to this basic approach. By incorporating stereo effects some

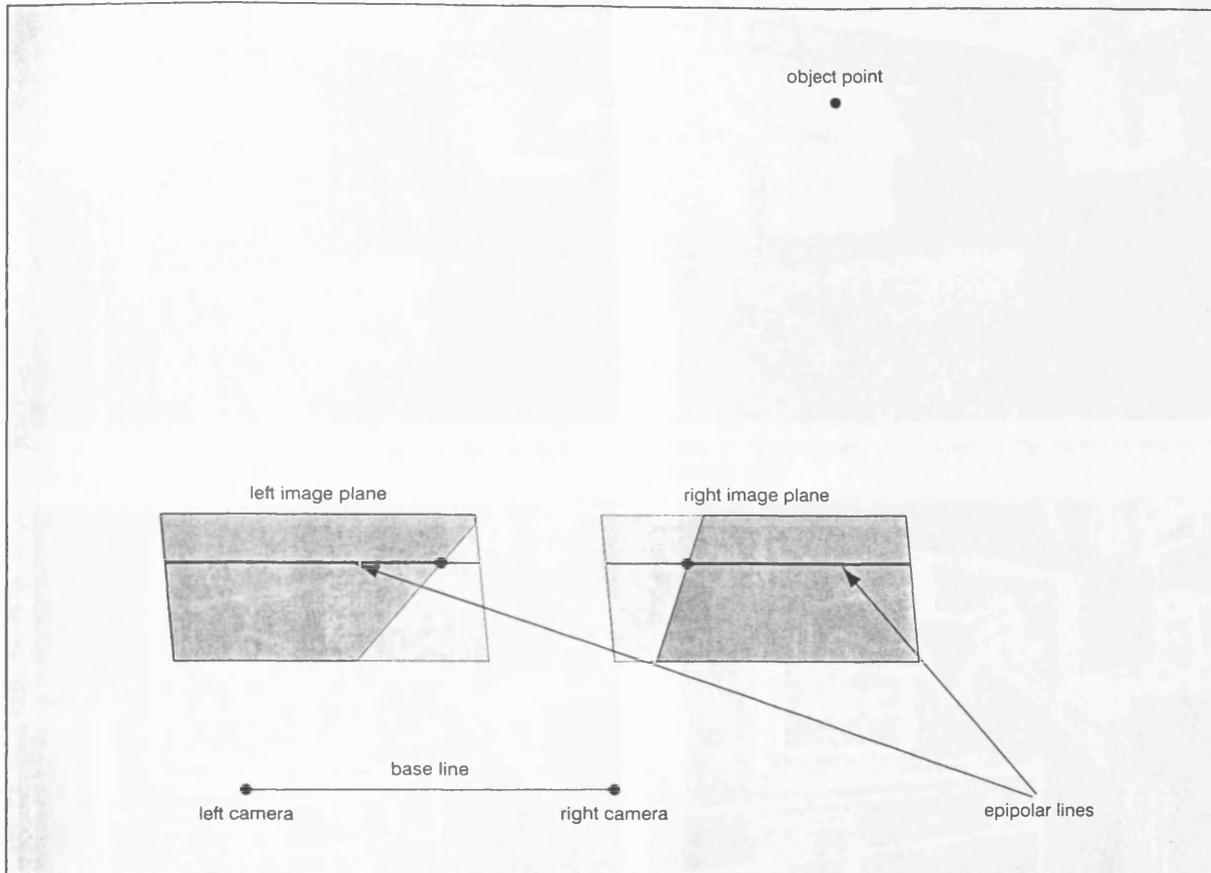


Fig. 5 Representation of the epipolar constraint for a pair of stereo cameras

idea of the location of various objects is given, thus improving the user's comprehension of the sound output. This has been achieved by simply assuming that the left side of the scene should generate sounds to be heard (mainly) in the left ear, and that the right side of the scene should generate sounds (mainly) in the right ear. Thus, the centre of the scene generates sounds that can be heard equally well in both ears. This representation immediately aids the subject in building a two-dimensional mental map of the scene from left to right.

Research by E. D. Schubert¹⁴ has shown encouraging results with the use of dichotic presentation for stereo effects. The human brain and auditory system is most adept at 'tuning in' on a particular sound source (i.e. a person), even when surrounded by a room full of talking people. One process that can be used to achieve this is through dichotic presentation. For example, if the sound source is directly in front of a person, then his or her brain expects the sound to reach both ears simultaneously. However, if the sound source is to the person's left, then the sound will be heard in the left ear a few milliseconds before it reaches the right ear. The two sounds are then fused together and used to reinforce each other. Techniques that incorporate this idea are still being developed.

Reduction of information content

Having taken the decision to include depth infor-

mation, the conventional techniques for data reduction in the optophone cannot be used. Loudness is normally associated with brightness, but a more natural representation is to associate loudness with proximity to the user. So, for example, an object that is close to the user would create a louder sound than an object far away. In order to achieve this the information about the colour or grey level of an object has to be discarded but the shape of the object preserved. This has been done by performing an edge detection routine on the original image, which produces cartoon-like images showing the outline of objects in the scene. The next step was to make the loudness of the corresponding sound output proportional to the proximity of the real-world edges.

Stereo disparity and depth maps

With two cameras it is possible to generate images that display the position/distance of objects according to their relative brightness. These images are known as depth maps. Although this process increases the computation time, it greatly decreases the information clutter in the image and, likewise, the chances of information overload for the user. Only objects that are of real significance (close) to the blind user are displayed. Objects further away have less impact on the user, and therefore have a lower amplitude/intensity in the final output.

A depth map gives the same information as

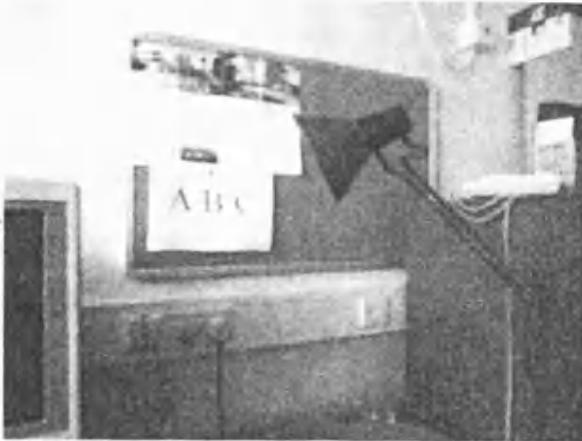


Fig. 6 Office scene as viewed by the left camera of a stereo pair



Fig. 7 Office scene as viewed by the right camera of a stereo pair

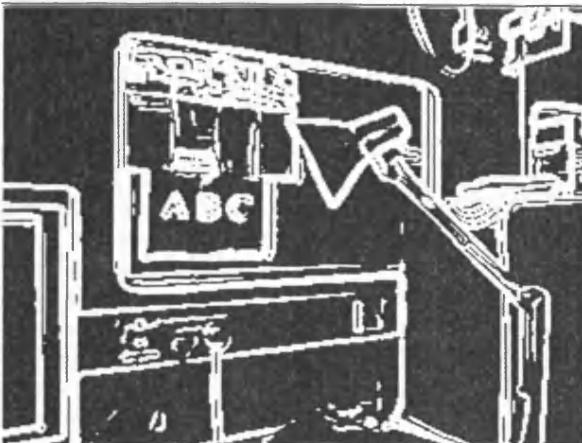


Fig. 8 Result of applying an edge detection algorithm to the left scene of Fig. 6

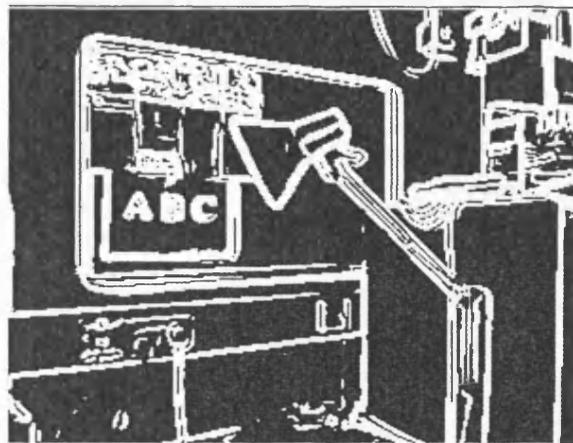


Fig. 9 Result of applying an edge detection algorithm to the right scene of Fig. 7

ultrasonic type devices, but not necessarily as fast or as effectively. Depth maps also have a second major drawback. Since the brightness corresponds to distance from the user, a page of text is perceived as a flat surface with uniform brightness. This means that the text on the page is lost so a depth map cannot be used in a reading device. This can be overcome with the use of an edge depth map.

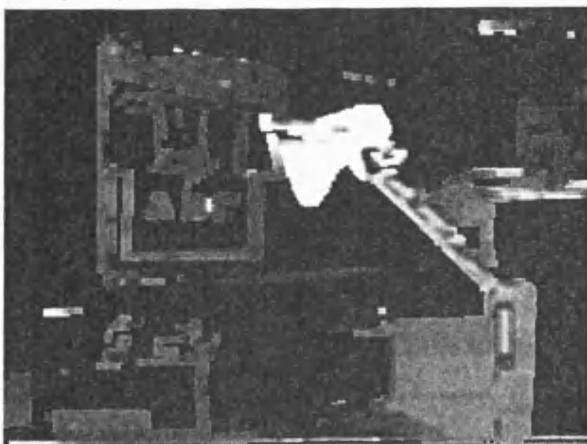


Fig. 10 Edge depth map after feature matching and applying a threshold

An edge depth map is a single image containing the edges of objects in a scene where the brightness of the edge corresponds to the proximity to the cameras. To obtain a depth map an edge detection routine is applied to images from two cameras and then a depth map created. Edge depth maps have the advantage of being able to produce outlines of objects and any text on a page while keeping all of the depth properties of a standard depth map.

Before a stereo edge depth map can be produced a method is needed to find the features in the two images that match. Two simple features that are used are the magnitude and orientation of the edges found in the images. Once the features have been found for the two camera images, and assuming they have been stored in image arrays, they are compared and a match found.

If the cameras are aligned correctly so that they lie along the same horizontal line, with the scan-lines for both cameras synchronised, then the so-called epipolar lines for both cameras will be along the corresponding image rows (see Fig. 5). This is known as the epipolar constraint, which means that a feature in one image (the left image say) has a matching feature in the corresponding row in the right image. As a result of this, the number of candidate matches, and hence the

computation time, is greatly reduced. Therefore, by using the epipolar constraint not only is it known on which row to search, but also in what direction to search.

Disparity is defined as the distance between the horizontal positions of a pair of matching features (if they could be overlaid onto the same image frame) and is measured in pixels. This is inversely proportional to depth, which is defined as the distance between the mid-point of the camera baseline and a real world object point¹⁵. So, when all features have been matched, the disparities can be converted into an intensity range or depth map.

Assuming there are n features to be matched with an upper disparity limit of d pixels, then without the epipolar constraint $n(2d)^2$ or $4n(d)^2$ pixels must be checked. However, using the epipolar constraint only nd pixels have to be compared.

Figs. 6 and 7 show two office scenes as viewed from two cameras set in stereo. In Figs. 8 and 9 they have been processed to reveal the strongest edges. The final image, Fig. 10, shows an actual stereo depth map as generated via a matching routine that is currently being developed. From this it can be seen that the shape of objects is preserved so that it is possible to identify and if necessary avoid nearby obstacles. Objects further away are still recognisable, but are much fainter. The information that is lost concerns the colour and some texture of the objects, which in this context is acceptable. However, any text in the scene can still be distinguished, as shown in Fig. 10.

The edge depth map is converted to stereo sound using the optophone transform described earlier. Because the scene contains edges only, there are large areas that do not create any sound at all. Hence the final sounds are much less cluttered and objects are easier to locate in these sounds.

Conclusions

The project has now been running for 18 months and the equipment and software are ready for testing. Both a Silicon Graphics O2 and a Pentium PC are being used with a 3-camera system. The software has been developed so that sounds can be generated from edge depth maps in real-time (up to a maximum of 15 frames per second). Over the next 18 months experiments will be conducted to determine the best way forward. The aim is to establish whether or not optophonic transformation can be made useful to blind users. However, there is a wider implication. If an optophone can generate sounds which in turn can generate a mental image that is clear enough to enable someone to 'see', does this have an application to the community as a whole? For example, sonification is an area where audio icons are being used as prompts in computing—a sound is generated which has a meaning to the user, such as 'you have e-mail'. Could the optophone be used to design icons that have both an easily recognisable

image and an easily identifiable sound that corresponds to that image? Recently a website showed how the optophone can be used in games in a system called the NANOVOICE 0.1—spectral sound synthesis on the Game Boy^{TM13}. This may be just the first of many applications in the entertainment industry.

Acknowledgment

The authors wish to thank the UK EPSRC for their funding under Research Grant Number GR/K85254.

References

- 1 MEIJER, P. B. L.: 'An experimental system for auditory image representations', *IEEE Trans. Biomed. Eng.*, February 1992, **BME-39**, (2), pp.112–121
- 2 MEIJER, P. B. L.: 'Image-audio transformation system.' US Patent No. 5,097,326, 1992
- 3 FOURNIER D'ALBE, E. E.: 'The moon element' (T. Fisher Unwin Ltd., London, 1924)
- 4 STARKIEWICZ, W., and KULISZEWSKI, T.: 'Active energy radiating system: the 80-channel Elektroftalm'. Proc. Int. Congress on Technology and Blindness, 1963, Vol. 1, pp.157–166 (The American Foundation for the Blind, New York)
- 5 COOPER, F. S. *et al.*: 'Evolution of reading machines for the blind: Haskin's Laboratories' research as a case history', *J. Rehabilitation Res. Devel.*, 1984, **21**, (1), pp.51–87
- 6 BEURLE, R. L.: 'Summary of suggestions on sensory devices'. St. Dunstan's, London, February 1947
- 7 KAY, L.: 'Electronic aids for blind persons: an interdisciplinary subject', *IEE Proc.*, 1984, **131**, Pt. A, (7), pp.559–576
- 8 FISH, R. M.: 'An audio display for the blind', *IEEE Trans. Biomed. Eng.*, March 1976, **BME-23**, (2), pp.144–154
- 9 DALLAS JR, S. A.: 'Sound pattern generator'. WIPO Patent Application No. WO 82/00395, 1982
- 10 O'HEA, A. R.: 'Optophone design: optical-to-auditory vision substitution for the blind'. PhD Thesis, The Open University, UK, April 1994
- 11 KURCZ, E.: 'Heliotrope: an optoelectronic aid for the blind', *Polish Tech. Rev.*, 1981, (4), pp.29–30
- 12 O'HEA, A. R.: 'A general-purpose optical-to-auditory seeing aid for the blind: design requirements and computer simulations'. MSc Dissertation, Brunel University Department of Computer Science, 1987
- 13 MEIJER, P. B. L. at http://ourworld.compuserve.com/homepages/Peter_Meijer
- 14 SCHUBERT, E. D.: 'Some preliminary experiments on binaural time delay and intelligibility', *J. Acoust. Soc. Am.*, September 1956, **28**, (5), pp.895–901
- 15 BERGENDAHL, J. R.: 'A computationally efficient stereo vision algorithm for adaptive cruise control'. MSc Thesis, Dept. Electrical Eng. & Computer Science, Massachusetts Institute of Technology, May/June 1997

© IEE: 2000

Phil Picton is Professor of Intelligent Computer Systems and Michael Capp is a postgraduate student in the School of Technology and Design, University College Northampton, St. George's Avenue, Northampton, NN2 6JD, UK

AN INVESTIGATION INTO STEREO VISION AS A MODIFICATION TO OPTOPHONIC MAPPINGS FROM SCENE-TO-SOUND

MICHAEL CAPP

PHIL PICTON

University College Northampton, UK

Abstract

This paper examines modern optophonic mapping techniques^{1,2} and discusses the role of stereo vision as an aid in the comprehension of the image-to-sound mapping. Techniques are also described for creating fast stereo depth maps, which present the user with only those parts of a real world scene considered essential for effective mobility. The outcome of this project was a real-time system, capable of portraying depth information to a user about the whereabouts of objects and obstacles in their immediate vicinity, whilst retaining the ability to perceive text.

Also mentioned are the key aspects of decision making (labelled DeLIA) involved when a blind subject encounters an obstacle, and results gained from preliminary tests designed to provide an insight into the performance of the optophonic process in these critical areas. The four aspects of blind mobility and navigation covered by DeLIA are: object detection, object location, object identification, and object avoidance.

Introduction

The modern day optophonic mapping, as invented by Peter Meijer^{1,3}, has been designed to enable conversion of images captured directly from a video camera into sound. This sound is encoded so that the location of an object in the real-world scene can be determined to a certain extent by the frequency (the higher the frequency, the higher the object) and the amplitude difference between the left and right ears (an object to the left would be portrayed with a greater sound amplitude in the left ear)^{3,4}. Similarly, a system known as dichotic presentation⁵ (incorporating minute delays between the sounds reaching the two ears) can be used to aid in portraying object location. However, the distance to an object is extremely hard to judge with any amount of accuracy. Secondly, with the current optophonic mapping there still appears to be a need to reduce the information contained within the captured image prior to conversion to sound due to the differing bandwidths between the human vision and auditory systems. For these reasons an investigation began into the role stereovision could play in a mobility aid for the blind.²

Information processing

Most stereo techniques tend to be very computationally intensive and at the start of the research real-time stereo algorithms were generally beyond the scope of standard desktop PC's. Hence, an

original stereo technique was derived that could run at speeds far beyond simple real-time (in excess of 150 frames/sec), generating an edge depth map where instead of distance being proportional to pixel intensity, it was related to the edge width⁶. After applying the optophonic scene-to-sound mapping the results were almost identical to that generated via a standard stereo edge technique. Since a thick edge produces a sound with a longer duration that is comparable to the high amplitude sound generated by a thin dark line, both of which are formed by nearby objects.

At this time one major drawback was encountered with this form of stereo depth map for the purposes at hand. When an edge depth map is generated for a particular scene it is often very demanding visually, to accurately determine individual objects in the original scene from the outlines alone. Thus, it was concluded that if a sighted person could not accurately determine the objects within a scene generated via a stereo edge algorithm, then how could a non-sighted person be expected to competently recognise the more demanding audio representation. A similar problem was also encountered by Pearson & Robinson⁷, and more recently by Manoranjan⁸, whilst studying the use of edge images for deaf hand signing over the telephone network via real-time low-resolution video. They found that subjects were unable to accurately interpret signing through the use of edge images alone. However, a suitable alternative was found in the form of two-tone cartoon edge images, which combined a simple edge image with a thresholded image.

By applying the cartooning technique to a stereo edge map it is possible to create a cartoon depth map⁹ (see Appendix - fig. (1)) that fills object regions, which allows successful recognition of the majority of objects in a scene. The algorithm employed simply calculates standard stereo edges, but as each stereo edge pixel is calculated, a subroutine proceeds to encode all neighbouring pixels to the same depth intensity whilst the condition holds that they belong to a filled area in the corresponding cartoon image. The advantages of this system are that objects regain definition, which is often lost during the application of an edge operator, and secondly that the whole process is capable of processing at a frame rate comparable to real-time video. Furthermore, stereo cartooning retains some texture from the original images, thus preserving the outline of any text that may be present.

In practice, stereo cartooning has been found to have an extra benefit over a standard intensity depth map for use as an input to the optophonic mapping, since it still consists of large regions that contain no information. These blank regions, which are relatively unimportant for basic mobility, generate no sound and thus, do not distract the user from more important details.

The conversion from scene-to-sound itself is quite an important factor in the optophonic mapping, and it is hoped that further advancements may be made to the current system by incorporating 3-D soundscapes to better portray the concepts of object location and distance to the user. Currently, as previously described, the sound output is generated in such a way as to mimic the natural functions of the human auditory system, with vertical height in the real world being logarithmically proportional to frequency^{3,4}. Object distance, as determined by the stereo algorithms, correspond to sound amplitude, such that nearer is louder. Each frame is then scanned from left to right at a rate of 1 or 2 frames per second.

Results

A series of tests have been devised, labelled DeLIA, to investigate four aspects of blind mobility and navigation believed to be important decision making stages in the detection and avoidance of obstacles. These four stages are: object Detection, object Location, object Identification, and object Avoidance.

Preliminary tests on a set of images with a group of sighted subjects have shown promising results using a real-time software driven optophone. These tests were designed to study the first three stages of DeLIA, that of Detection, Location and Identification, and shall be used as a guide for more extensive testing. Currently, the stereo cartooning technique has demonstrated an enhanced recognition rate for users in a shorter period of time than with previous optophonic systems using unmodified image inputs.

It was believed prudent to begin initial testing on sighted subjects alone since they tend to demonstrate an improved understanding of spatial awareness than a large percentage of the intended users of the optophone (the blind population). Thus, if the system had failed on trials with sighted subjects, then it would have been reasonable to assume that the advancements to the mapping would not have been successful for general use.

Acknowledgments

The authors wish to thank the UK EPSRC for their funding under Research Grant Number GR/K85254.

References

1. Meijer, P. B. L., http://ourworld.compuserve.com/homepages/Peter_Meijer. [Online] (Accessed March 2000).

2. Capp, M. and Picton, P., 'The Optophone: an Electronic Blind Aid', appearing in The Engineering Science and Education Journal, April 2000.
3. Meijer, P. B. L. (Feb. 1992). 'An experimental system for auditory image representations', *IEEE Trans Biomed Eng.*, vol. **BME-39**, no. 2, pp. 112-121.
4. Fish, R. M. (Mar. 1976). 'An audio display for the blind', *IEEE Trans Biomed Eng.*, vol. **BME-23**, no. 2, pp. 144-154.
5. Schubert, E. D. (Sep. 1956). 'Some preliminary experiments on binaural time delay and intelligibility.' *Journal of the Acoustical Society of America*, vol. **28**, No. 5, pp. 895-901.
6. Capp, M. and Picton, P., 'Fast, Low Resolution Edge Depth Maps and their Application to a Blind Mobility Aid', International Conference on Computer Vision, Pattern Recognition and Image Processing, Atlantic City, USA, February 27th – 3rd March 2000.
7. Pearson, D. E., & Robinson, J. A.: 'Visual communication at very low data rates.' *Proc. IEEE*, Apr. 1985, **73**(4), pp. 795-812.
8. Manoranjan, M.: 'Low bit rate communication using binary sketches for deaf sign language communication.' *M.Eng Thesis*, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, Aug. 1998, 159pp.
9. Capp, M. and Picton, P., 'Relaying Scene Information to the Blind via Sound using Cartoon Depth Maps.' Submitted to *IEE Proceedings – Vision, Image and Signal Processing*, 2000.

Appendix – Example Image

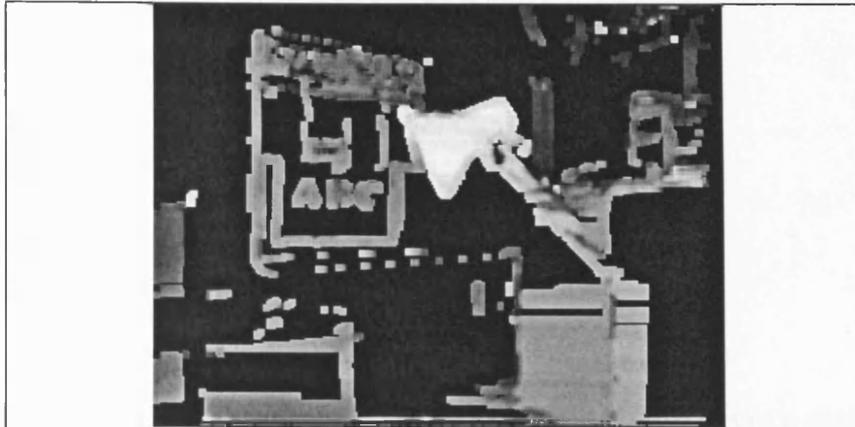


Figure (1) – Cartoon-edge depth map of an office scene. Closest objects are displayed in white for reasons of clarity.

Relaying Scene Information to the Blind via Sound using Cartoon Depth

Maps

Michael Capp

Prof. Phil Picton

Abstract

Important scene information can be relayed to a blind person via a mobility aid that incorporates an image to sound mapping. However, these devices are frequently plagued with being both tiring and stressful to use, whilst not forgetting their steep learning curve on behalf of the user. These difficulties are often a direct consequence of the quantity of information portrayed to the user when using a blind mobility aid that utilises near real-time scene-to-sound conversions. For this reason, research has been undertaken to consider and study the possibilities of stereo depth maps as a means to reducing the quantity of unnecessary information passed to the listener.

This paper describes, with examples, a method used for evaluating the effectiveness of edge depth maps. This is followed with an explanation of a new technique, which combines a stereo edge depth map with a cartoon-like scene representation, to effectively reduce the information content contained within a real-world image, whilst leaving the essential image details to allow for navigation and reading of large text.

Keywords

Real-time vision, stereoscopic vision, blind navigation aids.

1. Introduction

Optophonic transforms such as that invented by Peter Meijer^{3 4} for use with blind mobility aids are generally acknowledged as suitable scene-to-sound transforms and have been implemented in numerous devices⁴. However, blind mobility aids¹⁻⁸ that incorporate these scene-to-sound mappings, such as Peter Meijer's modern optophone^{3 4}, which is loosely based upon Fournier d'Albe's 'Exploring Optophone' and 'Reading Optophone'⁹, are still often plagued with the problem of overloading the user with information. In other words, the quantity of information they portray to the listener can rarely be accurately perceived for any length of time without considerable effort and concentration on behalf of the user. Hence, one possible remedy would be to reduce the quantity of information the mobility aid relays to the user. This process can therefore be tackled in two stages. The first stage is the reduction of information to produce a low-bandwidth image that contains as much relevant information as possible. The second stage is the conversion of the images to sound. In this paper a technique is demonstrated for stage one, which generates low-resolution stereo images. The standard optophonic transform is then used for stage two^{3 4} and is therefore not considered herein.

For stage one of the process, care must be taken when deciding what aspects of the real-world image can be discarded. As Adrian O'Hea⁶ pointed out, the scene-to-sound mapping should reflect the properties of the human vision system. For example, the human vision system is receptive to motion, changes in contrast, colours, and so on. Currently the optophonic mapping processes a monochrome image, captured from a web camera or similar device. The basic sound transform

mapping horizontal position along the image as a function of time, and vertical position with frequency. One aspect that seems to have been overlooked in previous optophonic mappings is the perception of depth. The human vision system allows us to judge the relative distances to real-world objects by processing the two images received through the eyes. This remarkable process can be emulated, albeit to a lesser standard, by using two or more video cameras¹⁰⁻¹⁴. At the expense of colour and shading information, it is possible to generate stereo depth maps that can be encoded with image pixel intensity being inversely proportional to object distance in the real world. Thus, the mapping from image-to-sound generates higher amplitude sounds for lighter image regions, which correspond to objects that are closer to the stereo cameras. This provides an opportunity to reduce the quantity of information portrayed to the user with the standard optophonic mappings by making distant objects very faint. However, this form of depth map often requires a great deal of processing time, and also lack the ability to portray text.

Further reductions in the quantity of information portrayed to the user can be made by utilising edge detection prior to generating the depth map¹⁵. The resulting edge depth map still encodes distance to real-world points by the relative pixel intensities, or alternatively via the thickness of the portrayed edges¹⁶. However, only the edges of objects in the depth image are visible, but this does allow for the portrayal of text.

There are a number of different methods that can be used to generate these stereo edge maps, as well as various techniques for performing the correlation procedure

itself. For example, one of the most common methods for identifying edge features to use for correlation, and the technique used in this paper, consists of passing the Sobel edge operator over the image frames. From this it is possible to obtain edge features such as the edge orientation, magnitude, and contrast changes. Then a decision must be taken over which correlation method will be used. Such as scanning for the first most likely match, or choosing a more time consuming method like finding all possible matches, then selecting the best and most likely correspondence.

When deciding on the best correlation procedure for generating edge depth maps, the norm tends to be to compare and evaluate a particular algorithm's performance in terms of its speed of operation and through visual inspection of the resulting edge depth map. Alternative methods are generally very time consuming, often involving the determination of a number of correspondences by hand¹¹. These selected matches are then compared with the results obtained from the stereo algorithm under trial and used as a basis on which to gauge its performance. The accuracy and reliability of this form of evaluation is then proportional to the sample of handpicked matches taken. Hence, a more suitable and automated method for determining the best procedure for detecting and generating an edge depth map was with the False Positive Fraction (FPF).

2. Evaluation of depth maps

The method used to evaluate the quality of a particular stereo edge algorithm requires a pair of stereo images in which the largest disparity in pixels is known, or can be determined. For example, the greatest disparity of any object that is

visible in both figures (2) & (3) (see Appendix), belongs to the head of the lamp, and is 30 pixels. Similarly, the smallest disparity between the two figures was found to be 15 pixels. Next the edge depth map is generated using the chosen algorithm. The disparity range used must be from the lowest disparity (in this case, 15 pixels) to one that exceeds the maximum disparity found between the stereo image pair (for example - 40 pixels). The algorithm must be applied so that the final edge depth map consists of edges with pixel intensities varying from black through to light grey (but not white). A pixel intensity of white is used to simply represent areas where no edge matches were located. Finally a histogram of pixel intensities is taken from the edge depth map.

From the histogram, all pixel intensities that correspond to matches that exceed a disparity of 30 pixels are known to be incorrect. Also, the total number of pixels corresponding to edges in the depth map is known. From these a ratio can be obtained of the total minus the incorrect matches to the total matches alone over the given range of disparities. Obviously, there are a number of assumptions in this procedure. For example, in most cases, as the maximum disparity increases, the number of false matches also increases. But, as long as all stereo algorithms under evaluation are tested using the same properties, a good indication of the algorithm's performance is obtained. A perfect algorithm would get a ratio of 1.

The following algorithm, when tested both visually and via the method described earlier for evaluating edge depth maps, gave some of the best results (see figure (6) in Appendix for an example) in the shortest computation time. The algorithm finds correspondences by searching for candidates between the image pair that

have near identical edge properties within a given disparity range, such as the edge magnitude, orientation, and contrast change. A second constraint that further reduces the number of incorrect matches is through the use of any previously found disparities¹⁵. It is often the case that several edge pixels are located from the same real world object. Hence, if two likely candidates are found, then priority is given to the one that most closely matches disparities that were correctly located on previous searches. Figure (1) shows an example where 'A' indicates a positive match between a pair of stereo images. Arrows 'B' & 'C' represent possible matches found while searching for the next correspondence. However, by considering the previous match ('A'), it can be assumed that match 'B' is incorrect and that match 'C' is valid and belongs to the same real world object as match 'A'.

Using the described method of evaluation on this stereo technique, with figures (2) & (3), the following results were obtained – see table (1). From these results it can be seen that only 4% of the matches are known to be incorrect which gives a ratio of $(12348 - 496)/12348 = 11852/12348 = 0.96$. This method has been applied to a number of image with a variety of content, texture, and illumination, and the FPF results obtained were accurate to within a few percent of the figure quoted.

As an aside, the original left and right edge images contained approximately 18200 pixels each that corresponded to edges. The edge-depth map contained 12348 pixels that correspond to edges, which is only 68% of the original edge pixels. This can be explained by the fact that the extreme left and right borders of

both images do not contain information that can be used for stereo matching. If these borders are excluded from the count, then the percentage of edges that appear in the edge-depth map is over 80%. The ratio obtained above is therefore a reasonably accurate measure of the algorithm's performance.

3. Cartooning Technique

An edge depth map is one solution to the problem of information reduction in a blind mobility aid. However, it can often be very difficult to interpret even relatively simple scenes when portrayed in the form of an edge depth map (see figure 6). For example, objects in the scene can become hard to recognise purely from their outlines, especially if parts of that outline are missing. This problem was also noted in a similar application by Pearson & Robinson¹⁷, on deaf sign communication over the phone network via real-time low-resolution video, and more recently by Manoranjan¹⁸. Due to the necessary bandwidth reductions, normal low-resolution video images occupied too much space in terms of data, and so could not be displayed at a sufficiently fast frame rate. On the other hand, two-tone edge images could be compressed quite adequately to allow for in excess of 8 frames per second, which was sufficiently fast enough to capture movement or signing. However, the loss of information due to edge detection made the accurate perception of signing very laborious and generally near impossible.

Their solution was to use two-tone cartoon images (see figure 7 for example) that contained sufficient information to allow for accurate interpretation of the hand signs, whilst suppressing enough unwanted scene information to relay the video at a comfortable frame rate for perception of movement.

The next method presented in this paper attempts to restore some of the lost texture (such as text) and shading into the final disparity image, by combining the edge depth map, discussed earlier, with the two-tone cartoon technique. The result is a cartoon-like edge depth map that greatly reduces the quantity of unnecessary information in the scene, whilst retaining object and region structure.

The cartoon image technique simply applied a threshold to an image. Any pixel intensities darker than the threshold were replaced with black pixels, all others were set to white. This, when combined with an edge detection routine, resulted in a two-tone cartoon filled image.

The threshold is determined using a method proposed by Manoranjan, whereby the absolute threshold was determined, via a histogram, using a fraction of the cumulative sum of pixel intensities in the image scene. Although this technique is not ideal, it does provide an adequate solution for most real world scenes.

The cartooning method was adapted to stereo images, so that edges and some surfaces would be displayed with a pixel intensity that was proportional to the distance from the cameras, as shown in figure (8). First, the left stereo image has the original cartoon technique applied, producing an image with edges and surfaces (see figure (7)). It was advantageous to set a third tone (other than black or white) in the cartoon that corresponds to object edges. This can be used to prevent region filling beyond object boundaries. Secondly, the right stereo image uses a Sobel edge detector to produce an edge image. A match is found between

an edge pixel in the left and right stereo images, and from that an estimate is made of the depth from the camera. The pixel in the edge-depth map is encoded with a brightness that is proportional to the distance from the camera as in the previous section. Using the cartoon image (the left stereo image) any surface that has been coded black can be re-coded using the same pixel intensity as the edge that has just been found, since it is assumed that the edge and the surface are at the same distance from the camera. This ceases when another match is found between the left and right images, which is assumed to be the opposite end of the surface.

4. Conclusion

This paper presents two methods for use with stereo depth maps. Firstly, a procedure for evaluating the effectiveness of stereo edge depth maps is illustrated, which can also be adapted for use in assessing the efficiency of standard intensity based depth maps. Secondly, this paper details a method for generating an edge depth map with a new and original presentation. An attempt is made to restore the areas of shading, which is lost whilst generating an edge depth map, by incorporating a two-tone cartoon image of the scene into the process. The resulting cartoon depth map represents the scene in the form of a cartoon, with objects being displayed in various shades of grey that represent the distance from the cameras.

The systems described are simple enough to work at speeds in excess of 15 frames per second, with adaptive cartoon thresholding. Once generated, this cartoon-depth map is converted into sound via a standard optophonic mapping, for interpretation by the listener/user. The method described successfully reduces the

quantity of relatively nonessential information in the scene, such as texture, whilst providing the user with depth and object information in the form of regions of like grey levels. This new method of presentation for stereo depth maps has already been shown, in preliminary experiments, to aid recognition and ease of interpretation of the optophonic output of a software implementation video optophone.

It should be noted that the technique for presenting images could make use of any stereo algorithm and so more accurate depth information could have been obtained using more up-to-date stereo algorithms. However, with technology at its current state, even a high specification PC would be unable to process the images, generate the depth maps, and perform the scene-to-sound mapping at sufficiently high speeds. With the stereo algorithm described in this paper, including the application of an original form of presentation (cartooning – as described earlier) and the inclusion of the scene-to-sound mapping, speeds of 12-15 frames per second were obtainable (image dimensions of 320x240).

The next step (stage 2) is the conversion to sound, and to more thoroughly test the effectiveness of the proposed techniques in trials with volunteers, results of which to be presented in Warwick, IMC10¹⁹, in August 2000.

5. Acknowledgments

The authors wish to thank the UK EPSRC for their funding under Research Grant Number GR/K85254.

6. References

1. BENHAM, T. A., & BENJAMIN, J. M., JR.: 'Active energy radiating systems: An electronic travel aid.' *Proceedings of the international congress on technology and blindness*, The American Foundation for the Blind: New York, (Editor – Clark, L. L), 1993, **1**, pp. 167-176.
2. KAY, L.: 'Electronic aids for blind persons: an interdisciplinary subject.' *IEE Review – IEE Proceedings*, 1984, vol. **131** (A-7), pp. 559-576.
3. MEIJER, P. B. L.: 'An experimental system for auditory image representations', *IEEE Trans Biomed Eng.*, Feb. 1992, **BME-39**(2), pp. 112-121.
4. MEIJER, P. B. L.: http://ourworld.compuserve.com/homepages/Peter_Meijer. [Online] (Accessed 2000).
5. NIELSON, L., MAHOWALD, M., & MEAD, C.: 'SeeHear.' *Proceedings: Image Analysis*, 5th Conference, Stockholm, 1987, **1**, pp. 383-396.
6. O'HEA, A. R.: 'Optophone Design: Optical-to-Auditory Vision Substitution for the Blind.' PhD Thesis, The Open University, UK, Apr. 1994.
7. SHOVAL, S., BORENSTEIN, J., & KOREN, Y.: 'The Navbelt – A Computerized Travel Aid for the Blind Based on Mobile Robotics Technology.' *IEEE Transactions on Biomedical Engineering*, Nov. 1998, **45**(11), pp. 1376-1386.

8. WARREN, D. H., & STRELOW, E. R., (Editors): 'Electronic spatial sensing for the blind.' *NATO Conference*, Lake Arrowhead, Sep. 1984, Martinus Nijhoff Publishers, 521pp.
9. FOURNIER D'ALBE, E. E.: 'The Moon Element – An introduction to the wonders of Selenium.' (T. Fisher Unwin Ltd. London: Adelphi Terrace, 1924).
10. BERGENDAHL, J. R.: 'A computationally efficient stereo vision algorithm for adaptive cruise control.' MSc thesis, Dep. of Electrical Engineering & Computer Science, Massachusetts Institute of Technology, 1997.
11. FUA, P.: 'A parallel stereo algorithm that produces dense depth maps and preserves image features.' *Machine Vision and Applications*, 1993, **6**, pp. 35-49.
12. MARSHALL, S., DURRANI, T. S., & CHAPMAN, R.: 'A binocular stereo matching algorithm.' *Computer & Control Colloquium on 'Vision Systems in Robotic and Industrial Control'*, The Institution of Electrical Engineers, Savoy Place, London, Mar. 1985.
13. MOLTON, N., SE, S., BRADY, J. M., LEE, D., & PROBERT, P.: 'A stereo vision-based aid for the visually impaired.' *Image and Vision Computing*, 1998, **16**, pp. 251-263.
14. MORAVEC, H. P.: 'Robot Spatial Perception by Stereoscopic Vision and 3D Evidence Grids.' CMU-RI-TR-96-34, The Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213, 1996, 44pp.
15. CAPP, M. & PICTON, P.: 'The Optophone: an Electronic Blind Aid', accepted for *The Engineering Science and Education Journal*, Apr. 2000.
16. CAPP, M. & PICTON, P.: 'Fast, Low-Resolution Edge Depth Maps and their Application to a Blind Mobility Aid.' *International Conference on Computer*

Vision, Pattern Recognition and Image Processing, JCIS2000, Atlantic City, Feb.-
Mar. 2000, **2**, pp. 248-251.

17. PEARSON, D. E., & ROBINSON, J. A.: 'Visual communication at very low data rates.' *Proc. IEEE*, Apr. 1985, **73**(4), pp. 795-812.

18. MANORANJAN, M.: 'Low bit rate communication using binary sketches for deaf sign language communication.' *M.Eng Thesis*, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John's, Newfoundland, Canada, Aug. 1998, 159pp.

19. CAPP, M. & PICTON, P. D.: 'An investigation into stereo vision as a modification to optophonic mappings from scene-to-sound.' Accepted for the 10th International Mobility Conference, Warwick, 4th-7th Aug. 2000.

7.1 Appendix – Tables

Disparity Range	Number of Pixels	Pixel Type
0	64452	Blank pixels
15-30	11852	Other matches
31-40	496	False matches
15-40	12348	Total matches

Table (1) – The results of applying the described method for evaluating stereo edge depth maps to the procedure presented in this paper, with figures (2) & (3).

7.2 Appendix – Figures

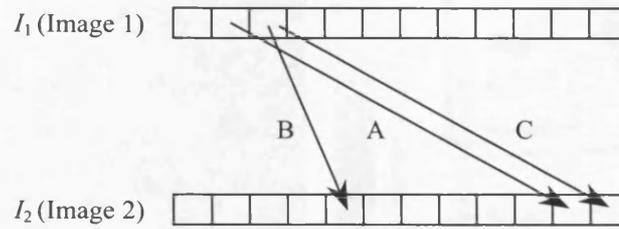


Figure (1) – Stereo matching

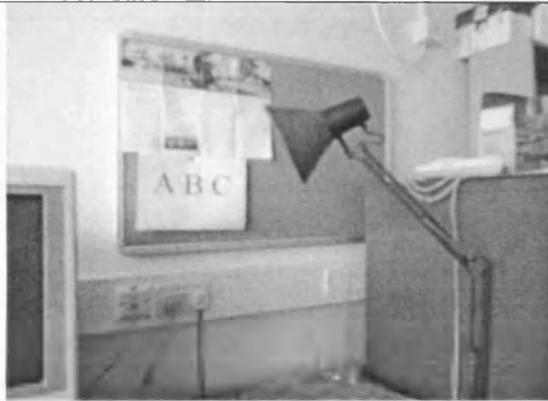


Figure (2) – Office scene captured through left camera of a stereo pair.

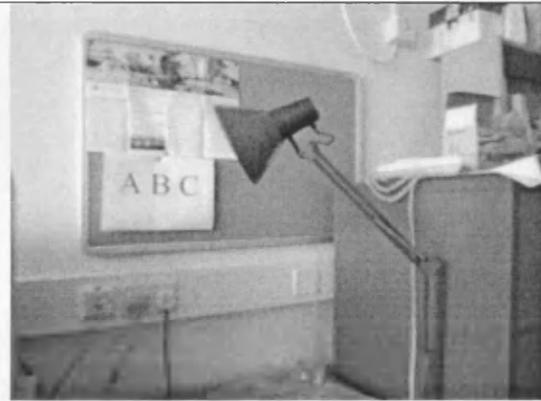


Figure (3) – Office scene captured through right camera of a stereo pair.

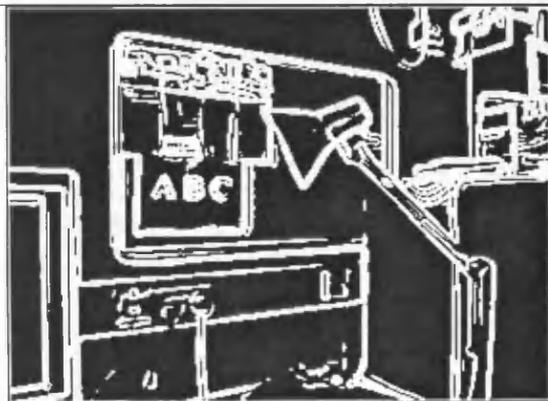


Figure (4) – Sobel edge operator applied to figure (2).

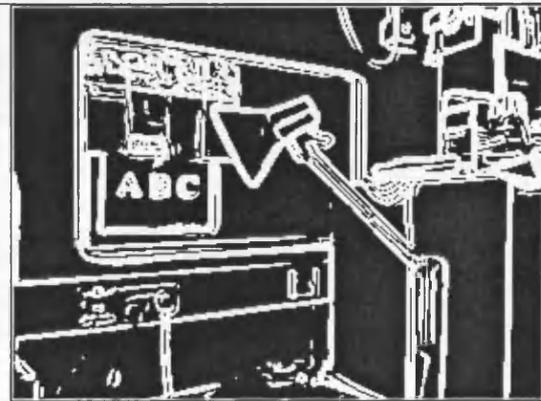


Figure (5) – Sobel edge operator applied to figure (3).

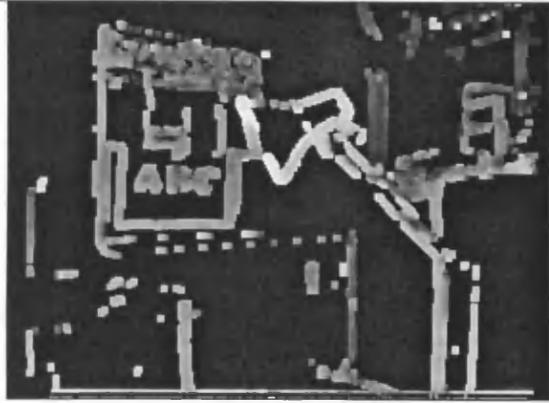


Figure (6) – Stereo edge depth map.

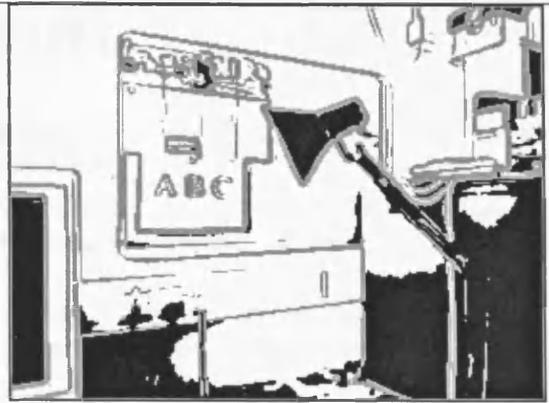


Figure (7) – Cartoon edge image.

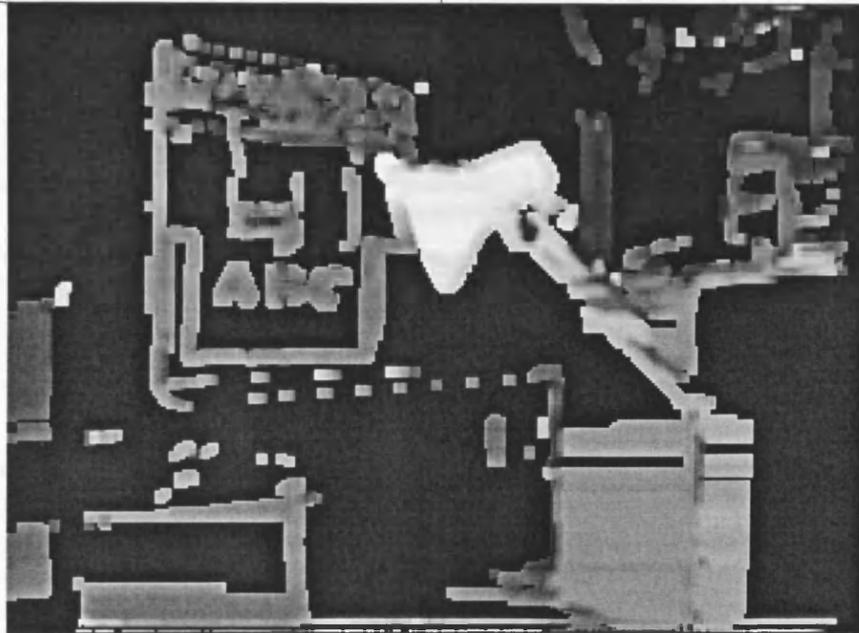


Figure (8) – Cartoon-edge depth map. Closest objects are displayed in white for reasons of clarity.

References

- [Arm96] – Armstrong, M. N. (1996). “Self-calibration from image sequences.” PhD Thesis, Department of Engineering Science, University of Oxford, UK.
- [BacColScaHolHar70] – Bach-Y-Rita, P., Collins, C. C., Scadden, L. A., Holmlund, G. W., & Hart, B. K. (1970). “Display technique in a tactile vision-substitution system.” *Medical and Biological Illustration*, vol. **20**, pp. 6-12.
- [BacHug85] – Bach-Y-Rita, P., & Hughes, B. (1985). “Tactile Vision Substitution: Some instrumentation and perceptual considerations.” In ‘*Electronic spatial sensing for the blind.*’ Editors Warren & Strelow, pp. 171-186.
- [BalBro82] – Ballard, D. H. & Brown, C. M. (1982). “Computer Vision.” *Prentice Hall*, ISBN 0-13-165316-4.
- [Bar21] – Barr, A., (Apr. 1921). “The Optophone.” *Journal of the Royal Society of Arts*, pp. 371-383.
- [Bar30] – Barnard, G. P., (1930). Chapter VIII – “The optophone: Photophony, or light-telephony.” Book entitled – *The Selenium Cell – Its properties and applications*, London – Constable & Company Ltd., pp. 240-259.
- [Bed68] – Beddoes, M. P. (Apr. 1968). “An inexpensive reading instrument with a sound output for the blind.” *IEEE Transactions on Biomedical Engineering*, vol. **BME-15**, No. 2, pp. 70-79.
- [BedSue71] – Beddoes, M. P., & Suen, C. Y. (Mar. 1971). “Evaluation and a method of presentation of the sound output from the Lexiphone – A reading machine for the blind.” *IEEE Transactions on bio-medical engineering*, vol. **BME-18**, no. 2, pp. 85-91.
- [BenAliSch73] – Benjamin, J. M., Ali, N. A., & Schepis, A. F. (1973). “A Laser Cane for the Blind.” *Proceedings of the San Diego Biomedical Symposium*, vol. **12**, pp. 53-57.
- [BenBen63] – Benham, T. A., & Benjamin, J. M., Jr., (1963). “Active energy radiating systems: An electronic travel aid.” *Proceedings of the international congress on technology and blindness*, The American Foundation for the Blind: New York, vol. **1**, (Editor – Clark, L. L), pp. 167-176.
- [Ber97] – Bergendahl, J. R. (May/June. 1997). “A computationally efficient stereo vision algorithm for adaptive cruise control.” *Dep. Of Electrical Engineering & Computer Science*, MSc Thesis, Massachusetts Institute of Technology, 56 pp.
- [Beu47] – Beurle, R. L. (Feb. 1947). “Summary of suggestions on Sensory Devices.” St. Dunstons, (JRS66), 4pp.
- [BorUlr97] – Borenstein, J. & Ulrich, I. (Apr. 1997). “The GuideCane – A Computerized Travel Aid for the Active Guidance of Blind Pedestrians.” *Proceedings of the IEEE International Conference on Robotics and Automation*, Albuquerque, NM, pp. 1283-1288.
- [Bra85] – Brabyn, J. (1985). “A review of mobility aids and means of assessment.” In Warren, D. H., & Strelow, E. R. (Eds.), ‘*Electronic Spatial Sensing for the Blind*’, 1985, 521pp. – pp. 13-27.
- [Bri98] – Brice, R. (Oct. 1998). “Improve your image.” *Electronics World - Audio*, October Edition, pp. 830-836.
- [BriGra87] – Brissaud, M., & Grange, G. (Nov. 1987). “Ultrasonic Sensory-Aid System for the Blind.” ‘*Systemes ultrasonores d’aide a la localisation pour non-voyants.*’ Research Notes, pp. 53-55
- [BurHanRis86] – Burns, J. B., Hanson, A. R., & Riseman, E. M. (Jul. 1986). “Extracting straight lines.” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. **PAMI-8**, no. 4, pp. 425-455.
- [CapPic00a] – Capp, M. and Picton, P. (Feb.-Mar. 2000). ‘Fast, Low Resolution Edge Depth Maps and their Application to a Blind Mobility Aid’, *International Conference on Computer Vision, Pattern Recognition and Image Processing*, Atlantic City, USA, **CVPRIP-12**, pp. 248-251.
- [CapPic00b] – Capp, M., & Picton, P. (Jun. 2000). “The Optophone: an Electronic Blind Aid.” *Engineering Science and Education Journal*, vol. **9**, no. 3, pp. 137-143.

- [CapPic00c] – Capp, M. & Picton, P., (Aug. 2000). ‘An investigation into stereo vision as a modification to optophonic mappings from scene-to-sound.’ *10th International Mobility Conference*, Warwick, UK, 4th-7th Aug. 2000.
- [CapPic00d] – Capp, M. & Picton, P., ‘Relaying Scene Information to the Blind via Sound using Cartoon Depth Maps.’ Submitted to *IEE Proceedings – Vision, Image and Signal Processing*, 2000.
- [ClaHeyHow86] – Clark-Carter, D. D., Heyes, A. D., & Howarth, C. I. (1986). “The effect of non-visual preview upon the walking speed of visually impaired people.” *Ergonomics*, vol. **29**, no. 12, pp. 1575-1581.
- [Cle98] – Clegg, F. (1998). “Simple statistics – A course book for the social sciences.” Cambridge University Press, Fifteenth Printing 1998, ISBN 0 521 28802 9.
- [Cof63] – Coffey, J. L. (1963). “The development and evaluation of the Battelle aural reading device.” *Proceedings of the International Congress on Technology & Blindness*, The American Foundation for the Blind: New York, vol. **1** of 4, pp. 343-360.
- [Col85] – Collins, C. C. (1985). “On mobility aids for the blind.” In Warren, D. H., & Strelow, E. R. (Eds.), ‘*Electronic Spatial Sensing for the Blind*’, 1985, 521pp. – pp. 35-64.
- [CooGaiNye84] – Cooper, F. S., Gaitenby, J. H., & Nye, P. W. (1984). “Evolution of Reading machines for the Blind: Haskin’s Laboratories’ Research as a Case History.” *Journal of Rehabilitation Research and Development*, vol. **21**, 1, pp. 51-87.
- [DalEri80] – Dallas Jr, S. A., & Erickson, A. J. (1980). “Sound pattern generator”, World Intellectual Property Organisation patent application no. WO 82/00395.
- [Dee85] – Deering, M. F. (1985). “Computer vision requirements in blind mobility aids.” In Warren, D. H., & Strelow, E. R. (Eds.), ‘*Electronic Spatial Sensing for the Blind*’, 1985, 521pp. – pp. 65-82.
- [DesRag90] – Desu & Raghavarao (1990), pp. 3-4.
- [Dew99] – Dewhurst, D. (Accessed April 1999). “Vuphonics.” <http://ourworld.compuserve.com/homepages/DavidDewhurst/Vuphonics.html>. [Online].
- [Dod85] – Dodds, A. G. (1985) – “Evaluating mobility aids: An evolving methodology.” In ‘*Electronic spatial sensing for the blind.*’ Editors Warren & Strelow, pp. 192-200.
- [DudHar72] – Duda, R. O., & Hart, P. E. (Jan. 1972). “Use of the Hough Transformation to Detect Lines and Curves in Pictures.” *Communications of the ACM*, vol. **15**, No, 1, pp. 11-15.
- [EleRev21] – (Feb. 1921). “The optophone – An instrument to enable the Blind to read.” *The Electrical Review*, vol. **88**, No. 2255, pp. 166-168.
- [Fis73] – Fish, R. M. (Oct. 1973). “A sensory aid for the blind using an audio display.” *26th ACEMB*, p. 144.
- [Fis76] – Fish, R. M. (Mar. 1976). “An audio display for the blind”, *IEEE Trans Biomed Eng.*, vol. **BME-23**, no. 2, pp. 144-154.
- [Fou14] – Fournier d’Albe, E. E. (Jun. 1914). “A type reading Optophone.” *Nature*, vol. **93**, No. 2328, p. 394.
- [Fou24] – Fournier d’Albe, E. E. (1924). “The Moon Element”, Book, T. Fisher Unwin Ltd, London.
- [Fos64] – Foster, D. B. (Aug. 1964). “A system for use by the blind for environment identification.” London Patent Office – Patent Specification 1,106,671. Date of filing: 26th Aug. 1964. Date Published: 20th Mar. 1968.
- [FriBacTomWeb87] – Frisken-Gibson, S. F., Bach-Y-Rita, P., Tompkins, W. J., & Webster, J. G. (Dec. 1987). “A 64-Solenoid, four-level fingertip search display for the blind.” *IEEE Transactions on Biomedical Engineering*, vol. **BME-34**, no. 12, pp. 963-965.
- [Fua93] – Fua, P., (1993). “A parallel stereo algorithm that produces dense depth maps and preserves image features.” *Machine Vision and Applications*, vol. **6**, pp. 35-49.
- [GeiKla94] – Geilser, W. S., & Klarquist, W. N. (Dec. 1994). “Maximum-Likelihood Method for Image Deblurring and Depth-from-Defocus.” *Technical Report – UT-CVIS-TR-94-008*, Center for Vision and Image Sciences, The University of Texas at Austin, 25pp.

- [GeiPer98] – Geisler, W. S., & Perry, J. S. (1998). “A real-time foveated multiresolution system for low-bandwidth video communication.” Obtained from web site: ‘<http://fi.cvis.psy.utexas.edu/publicat.htm>’, accessed April 2000, 13pp., University of Texas Centre for Vision and Image Sciences, Austin, Texas 78712.
- [GeiPer99] – Geisler, W. S., & Perry, J. S. (1999). “Variable-Resolution Displays for Visual Communication and Simulation.” Obtained from web site: ‘<http://fi.cvis.psy.utexas.edu/publicat.htm>’, accessed April 2000, 4pp., University of Texas Centre for Vision and Image Sciences, Austin, Texas 78712.
- [GenMor76] – Gennery, D., & Moravec, H. (Sep. 1976). “Cart Project Progress Report.” Stanford University.
- [HarSte88] – Harris, C., & Stephens, M. (1988). “A combined corner and edge detector.” *Fourth Alvey Vision Conference*, pp. 147-151.
- [HayJai83] – Haynes, S. M., & Jain, R. (1983). “Detection of Moving Edges.” *Computer Vision, Graphics, and Image Processing*, vol. **21**, pp. 345-367.
- [Hey85] – Heyes, A. D. (1985). “Microprocessor techniques applied to ultrasonic pulse/echo travel aids for the Blind.” In Warren, D. H., & Strelow, E. R. (Eds.), ‘*Electronic Spatial Sensing for the Blind*’, 1985, 521pp. – pp. 161-169.
- [Hil82] – Hilfreth, E. C. (Aug. 1982). “Edge detection for computer vision system.” *Mechanical Engineering*, pp. 48-53.
- [HilDodHilFox95] – Hill, M-M., Dodson-Burk, B., Hill, E. W., & Fox, J. (Jul/Aug. 1995). “An infant sonicguide intervention program for a child with a visual disability.” *Journal of Visual Impairment & Blindness*, vol. **89**, pp. 329-336.
- [Hol94] – Hollander, A. J. (1994). “An exploration of virtual shape perception.” *Engineering MSc*, University of Washington.
- [IfuSasPen91] – Ifukube, T., Sasaki, T., & Peng, C. (May 1991). “A Blind mobility aid modeled after echolocation of bats.” *IEEE Transactions on Biomedical Engineering*, vol. **38**, No. 5, pp. 461-465.
- [Jac63a] – Jacobson, B. (1963). “Passive Systems: Ambient-light obstacle detector with tactile output.” *Proceedings of the international congress on technology & blindness*, The American Foundation for the Blind: New York, vol. **1** of 4, no. 5, pp. 187-192.
- [Jac63b] – Jacobson, B. (1963). “Passive Systems: A magnetic compass and straight course indicator for the blind.” *Proceedings of the international congress on technology & blindness*, The American Foundation for the Blind: New York, vol. **1** of 4, no. 5, pp. 193-197.
- [JaiKasSch95] – Jain, R., Kasturi, R., & Schunck, B. G. (1995). “Machine Vision.” *McGraw-Hill International Editions*, Computer Science Series, ISBN 0-07-113407-7.
- [Joh63] – Johnson, A. R. (1963). “Passive Systems: A proposed stereo-optical edge detector.” *Proceedings of the international congress on technology & blindness*, The American Foundation for the Blind: New York, vol. **1** of 4, no. 5, pp. 183-186.
- [KacWebBacTom91] – Kaczmarek, K. A., Webster, J. G., Bach-Y-Rita, P., & Tompkins, W. J. (Jan. 1991). “Electrotactile and vibrotactile displays for sensory substitution systems.” *IEEE Transaction on Biomedical Engineering*, vol. **38**, no. 1, pp. 1-16.
- [Kay84] – Kay, L. (1984). “Electronic aids for blind persons: an interdisciplinary subject.” *IEE Proc*, vol. **131**, part A, no. 7, pp. 559-576.
- [Kay85] – Kay, L. (1985). “Sensory aids to spatial perception for blind persons: Their design and evaluation.” In ‘*Electronic Spatial Sensing for the Blind*.’ Editors Warren & Strelow, 521pp. - pp. 125-139.
- [KenCok92] – Kent, M., & Coker, J. (1992). “Vegetation description and analysis – Practical approach.” *John Wiley & Sons*, ISBN 0 471 94810 1.
- [KlaGeiBov95] – Klarquist, W. N., Geisler, W. S., & Bovik, A. C. (Aug. 1995). “Maximum-Likelihood Depth-from-Defocus for Active Vision.” *IEEE International Conference on Intelligent Robots and Systems*, Pittsburgh, Pennsylvania, 6pp.

- [KorZim86] – Kories, R., & Zimmermann, G. (1986). “A Versatile Method for the Estimation of Displacement Vector Fields from Image Sequences.” *IEEE*, CH2322-6/86/0000/0101, pp. 101-106.
- [KorGei96] – Kortum, P., & Geisler, W. (1996). “Implementation of a foveated image coding system for image bandwidth reduction.” Obtained from web site: ‘<http://fi.cvis.psy.utexas.edu/publicat.htm>’, accessed April 2000, 13pp., University of Texas Centre for Vision and Image Sciences, Austin, Texas 78712.
- [Kur81] – Kurcz, E. (1981). “Heliotrope: an optoelectronic aid for the blind.” *Polish Tech Rev*, no. 4, pp. 29-30.
- [Lau68] – Lauer, H. L. (1968). “The Visotoner: A personal reading machine for the blind.” *Bulletin Prosthetics Research*, vol. **10/9**, pp. 99-103.
- [Lau89] – Lauer, H. (1989). “Reading machines for the blind: Why one medium isn’t enough.” *RDC Newsletter*, Jan.-Feb. 1989, pp. 10-15.
- [Lip58] – Lipton, A. (Jun. 1958). “The development and evaluation of aural reading devices for the blind – *Optophone-type reading machine for the blind*.” Abstract based upon final report, Veterans Administration by Battelle Memorial Institute, Columbus, Ohio, June 30, 1958.
- [Low91] – Low, A. (1991). “Introductory Computer Vision and Image Processing.” *McGraw-Hill*, ISBN 0-07-707403-3.
- [Mac85] – Mackay, R. S. (1985). “Physical principles underlying blind guidance prostheses with an emphasis on the ultrasonic exploration of a region of space.” In Warren, D. H., & Strelow, E. R. (Eds.), ‘*Electronic Spatial Sensing for the Blind*’, 1985, 521pp. – pp. 141-153.
- [Man98] – Manoranjan, M. (Aug 1998). “Low bit rate communication using binary sketches for deaf sign language communication.” *A thesis submitted to the school of graduate studies in partial fulfilment of the requirements for the degree of master of engineering*, Faculty of Engineering and Applied Science, Memorial University of Newfoundland, St. John’s, Newfoundland, Canada. 159pp.
- [MarDurCha85] – Marshall, S., Durrani, T. S., & Chapman, R., (March 1985). “A binocular stereo matching algorithm.” *Computer & Control Colloquium on ‘Vision Systems in Robotic and Industrial Control’*, The Institution of Electrical Engineers, Savoy Place, London.
- [MarHil80] – Marr, D., & Hildreth, E. (1980). “Theory of edge detection.” *Proc. R. Soc. Lond.*, vol. B **207**, pp. 187-217.
- [MciMut88] – McIntosh, J. H., & Mutch, K. M. (1988). “Matching straight lines.” *Computer Vision, Graphics, and Image Processing*, vol. **43**, pp. 386-408.
- [Mei92a] – Meijer, P. B. L. (Feb. 1992). “An experimental system for auditory image representations”, *IEEE Trans Biomed Eng.*, vol. **BME-39**, no. 2, pp. 112-121.
- [Mei92b] – Meijer, P. B. L. (1992) “Image-Audio Transformation System”, United States Patent No. 5,097,326.
- [Mei00] – Meijer, P. B. L. (Accessed May 2000). http://ourworld.compuserve.com/homepages/Peter_Meijer. [Online].
- [MetPan99] – Metz, C. E., & Pan, X. (1999). “ ‘Proper’ binormal ROC curves: Theory and maximum-likelihood estimation.” *Journal of Mathematical Psychology*, vol. **43**, pp. 1-33.
- [Moo86] – Moore, G. (Jul. 1986). “Literacy and computing for the blind.” *Electronics & Power*, pp. 513-516.
- [Mor96] – Moravec, H. P. (1996). “Robot Spatial Perception by Stereoscopic Vision and 3D Evidence Grids.” Technical Report CMU-RI-TR-96-34, The Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. 44pp.
- [NieMahMea87] – Nielsen, L., Mahowald, M. A., & Mead C. (1987). “SeeHear.” *Proceedings: Image Analysis, 5th Conference*, Stockholm, vol. **1**, pp. 383-396.
- [Nye70] – Nye, P. W. (Apr. 1970). “Human factors underlying the design of reading aids for the blind.” *IEEE Transactions on Biomedical Engineering*, vol. **BME-17**, no. 2, pp. 97-100.

- [Ohe87] – O’Hea, A. R. (1987). “A general-purpose optical-to-auditory seeing aid for the blind: design requirements and computer simulations”, MSc Dissertation, Brunel University Department of Computer Science, 105 pp.
- [Ohe94] – O’Hea, A. R. (Apr. 1994). “Optophone Design: Optical-to-Auditory Vision Substitution for the Blind.” PhD Thesis, The Open University, UK.
- [PeaRob85] – Pearson, D. E., & Robinson, J. A. (Apr. 1985). “Visual communication at very low data rates.” *Proc. IEEE*, vol. 73, no. 4, pp. 795-812.
- [Pic84] – Picton, P. D. (Mar. 1984). “Hough Transforms.” Dept. of Electrical & Electronic Engineering, Heriot-Watt University, Edinburgh, pp. 1-16.
- [Pic89] – Picton, P. D. (Jul 1989). “Tracking and segmentation of moving objects in a scene.” *IEE 3rd International Conference on Image Processing and its Applications*, Conference Publication No. 307, pp. 389-393.
- [PolMayFri85] – Pollard, S. B., Mayhew, J. E. W., & Frisby, J. P. (1985). “PMF: A stereo correspondence algorithm using a disparity gradient limit.” *Perception*, AI Vision Research Unit, University of Sheffield, vol. 14, pp. 449-470.
- [PGR00] – Point Grey Research (Accessed April 2000). <http://www.ptgrey.com>. [Online]
- [RajCha97] – Rajagoplan, A. N., & Chaudhuri, S. (Dec. 1997). “Space-Variant Approaches to Recovery of Depth from Defocused Images.” *Computer Vision and Image Understanding*, vol. 68, no. 3, pp. 309-329.
- [Row70] – Rowel, D. (1970). “Auditory factors in the design of a binaural sensory aid for the blind.” PhD Thesis, University of Canterbury, Canterbury, New Zealand.
– Cited by [ShoBorKor98].
- [Sch56] – Schubert, E. D. (Sep. 1956). “Some preliminary experiments on binaural time delay and intelligibility.” *Journal of the Acoustical Society of America*, vol. 28, No. 5, pp. 895-901.
- [SciAme20] – (Nov. 1920). “The Type-Reading Optophone – An instrument which enables the Blind to read ordinary type.” *Scientific American*, vol. CXXIII, No. 19, p. 463.
- [ShoBorKor98] – Shoval, S., Borenstein, J., & Koren, Y. (Nov. 1998). “The Navbelt – A Computerized Travel Aid for the Blind Based on Mobile Robotics Technology.” *IEEE Transactions on Biomedical Engineering*, vol. 45, no. 11, pp. 1376-1386.
- [SkuConGor99] – Skudlarski, P., Constable, R. T., & Gore, J. C. (1999). “ROC analysis of statistical methods used in functional MRI: Individual subjects.” *NeuroImage*, vol. 9, pp. 311-329.
- [Smi77] – Smith, K. (Feb. 1977). “Seeing with microcircuits.” *New Scientist*, 3rd Feb. 1977, p.260.
- [Spu85] – Spungin, S. J. (1985). “Technology and the Blind person: Corridors of insensitivity.” In Warren, D. H., & Strelow, E. R. (Eds.), ‘*Electronic Spatial Sensing for the Blind*’, 1985, 521pp. – pp. 375-386.
- [StaKul63] – Starkiewicz, W., & Kuliszewski, T. (1963). “Active Energy Radiating System: The 80-Channel Elektroftalm.” *Proc. Of the International Congress on Technology and Blindness*, The American Foundation for the Blind: New York, vol. 1 of 4, pp. 157-166.
- [TouAdj85] – Tou, J. T. & Adjouadi, M. (1985). “Computer vision for the blind.” In Warren, D. H., & Strelow, E. R. (Eds.), ‘*Electronic Spatial Sensing for the Blind*’, 1985, 521pp. – pp. 83-124.
- [Tur02] – Turine, V. de. (1902). “Photophonic books for the blind.” *L’Eclairage Electrique*, vol. 31, pp. 16-19.
- [VerWan85] – Veraart, C., & Waner, M. –C. (1985). “Sensory substitution of vision by audition.” In ‘*Electronic spatial sensing for the blind*,’ Editors Warren & Strelow, pp. 217-238.
- [Wal85] – Wallace, R. S. (1985). “A Modified Hough Transform for Lines.” *Proceedings IEEE Computer Vision and Pattern Recognition*, pp. 665-667. ISBN 0 81 8606339.
- [WarStr85] – Warren, D. H., & Strelow, E. R., (Editors) (1985). “Electronic spatial sensing for the blind.” ISBN: 90-247-3238-7, 521pp.