Single-imager occupant detection based on surface reconstruction

Submitted for the Degree of
Doctor of Philosophy
at
City University
London

By
Jason JeongSuk Yoon

December 28, 2004

Contents

Abstract XVII
Declaration XIX
Acknowledgement XXI

1 Introduction **1**
1.1 Aim of this work 2
1.2 Vision-based occupant detection systems 3
1.2.1 Motivation 3
1.2.2 Occupant detection systems 5
1.2.3 State of the art 7
1.3 System overview 11
1.4 Organisation of the thesis 13

2 Acquisition and pre-processing **15**
2.1 Motivation 16
2.2 Sensors and illuminations 17
2.2.1 Optical dynamic range 17
2.2.2 Imaging sensors 18
2.2.3 Active illumination 24
2.3 Image enhancement: DoubleFlash 24
2.3.1 Introduction 24
2.3.2 Offset reduction 26
2.3.3 Dynamic range compression 27
2.3.4 Experimental results 28
2.4 Shadow removal: ShadowFlash 28
2.4.1 Motivation 28
2.4.2 Analysis 30
2.4.3 Shadow removal 31
2.4.4 Experimental results 38
2.4.5 Discussion 41
2.5 Precis 42

3 2D processing: object segmentation **45**
3.1 Motivation . . . 46
3.1.1 Introduction . . . 46
3.1.2 Segmentation for a vehicle cabin environment . . . 48
3.1.3 Overview . . . 49
3.2 Approximate boundary extraction . . . 50
3.2.1 Texture-based object detection . . . 50
3.2.2 Morphological operations . . . 52
3.2.3 Deciphering of object of interest . . . 53
3.3 Active contour models . . . 53
3.3.1 Introduction . . . 53
3.3.2 Fundamentals . . . 54
3.3.3 Dynamic programming . . . 57
3.3.4 Convexity defects driven active contour models . . . 60
3.4 Experimental results . . . 61
3.5 Precis . . . 63

4 3D processing: surface reconstruction **73**
4.1 Motivation . . . 74
4.2 3D sensing techniques and their limitations . . . 74
4.2.1 Ultrasonic imaging . . . 75
4.2.2 Laser scanning . . . 75
4.2.3 Structured lighting . . . 76
4.2.4 Time-of-flight cameras . . . 76
4.2.5 Shape from shading . . . 76
4.2.6 Stereo vision . . . 77
4.3 Stereo vision techniques . . . 78
4.3.1 Fundamentals . . . 78
4.3.2 Limits and drawbacks . . . 80
4.4 Photometric Stereo Method . . . 83
4.4.1 Introduction . . . 83
4.4.2 Reflection models . . . 84
4.4.3 Surface normal estimation . . . 85
4.4.4 Surface integration . . . 87
4.4.5 Advantages and drawbacks . . . 89
4.5 Experimental Results . . . 92
4.5.1 Reconstruction examples . . . 92
4.5.2 Evaluation of the reconstructed surface accuracy . . . 93
4.5.3 Surface reconstruction of temporal sequences . . . 93
4.6 Precis . . . 94

5 Classification **101**
5.1 Introduction . . . 102
5.1.1 Motivation . . . 102
5.1.2 State of the art . . . 102
5.1.3 Summary . . . 104
5.2 Feature selection . . . 104
5.2.1 Extended Gaussian image . . . 104
5.2.2 Surface depth . . . 105
5.2.3 Spread axes information . . . 105
5.2.4 Relative position of the upper extremum . . . 107
5.2.5 Volumetric ratio and compactness . . . 107
5.2.6 Other 2D geometric information . . . 108
5.3 Classifier design . . . 108
5.3.1 Occupant class assignment and operation assumption 108
5.3.2 System design requirements . . . 110
5.3.3 Neural networks . . . 110
5.3.4 Implementation . . . 114
5.4 Precis . . . 116

6 Experimental results **119**
6.1 Introduction . . . 120
6.2 Experimental setup . . . 120
6.2.1 Algorithm implementation and hardware embodiment 120
6.2.2 Data collection . . . 122
6.3 Evaluation . . . 123
6.3.1 Processing time . . . 123
6.3.2 Feature consistency . . . 124
6.3.3 Network training . . . 124
6.3.4 Classification performance evaluation . . . 125
6.4 Discussion . . . 127

7 Conclusions **133**
7.1 Precis . . . 134
7.2 Assessment . . . 135
7.3 Contribution . . . 136
7.4 Future work . . . 137

Bibliography . . . 139

List of Tables

1.1 Description of the predefined occupant classes 5

6.1 Processing time consumed in each processing module 124
6.2 Error statistics *without* the tapped delay lines. Overall error rate: 6.66% 126
6.3 Error statistics *with* the tapped delay lines. Overall error rate: 1.14% 127
6.4 Error statistics of the system employing the weighted averaging approach 127

List of Figures

1.1 The statistics of the fatalities 4
1.2 Simulation of an unbelted child during an airbag deployment 5
1.3 Illustrations for the six occupant classes 6
1.4 System overview 12

2.1 Typical examples of image detail lost 17
2.2 Irradiance variations during motorway drive 19
2.3 Irradiance variations in a parking lot 19
2.4 Drive through the city of Munich including tunnels 20
2.5 Comparison between different response functions of CMOS cameras 22
2.6 SollyCam version 3.0 23
2.7 Spectral irradiance of the sunlight after atmosphere 25
2.8 Transfer function of a typical NIR bandpass filter 25
2.9 The result of the DoubleFlash method 28
2.10 Illustration of penumbra and umbra within a shadow scene caused by a spot light source 30
2.11 Penumbrae problem 31
2.12 Shadows on an overcast day 32
2.13 Illustration for the formation of shadows with two spot light sources 33
2.14 A Venn diagram based on the amount of the irradiance power 34
2.15 Illustration of the shadow removal procedure 35
2.16 Comparison between the non-sliding and sliding N-tuple strategy 37
2.17 Examples of ShadowFlash with the ambient illumination . . . 39
2.18 Intensity histograms for the inputs and the result where the ambient light exists 40
2.19 Examples of ShadowFlash for colour images taken by a CCD camera with AGC 41
2.20 Examples of ShadowFlash for outdoor images 41
2.21 Less successful case due to the uneven distributed illumination 42
2.22 Sample sequence of Real-time ShadowFlash 43

3.1 Illustration for the deformation of an active contour model . . 56
3.2 Illustration for the movement of a snaxel with respect to the continuity energy 57
3.3 Illustration of the snaxel movement according to the balloon force 58
3.4 Demonstration for a snake using dynamic programming . . . 59
3.5 The definition of convexity defects 60
3.6 Typical examples for an active contour model with support of convexity defects 61
3.7 The reference background used in the experiment 62
3.8 Segmentation result applied to the *Adult* class 65
3.9 Segmentation result applied to the *FFCS* class 66
3.10 Segmentation result applied to the *RFCS* class 67
3.11 The effect of the ShadowFlash technique applied to the segmentation process 68
3.12 Another example for segmentation via the ShadowFlash technique 69
3.13 The evaluation of segmentation result with the ShadowFlash technique 70
3.14 Example of less successful segmentation 71

4.1 Non-standard stereo geometry vs. standard stereo geometry . 78
4.2 Illustration of the basic concept of the stereo vision technique 79
4.3 The result of a poor correspondence analysis 80
4.4 Illustration of the set of solutions of the different shading based shape recovery methods 86
4.5 Illustration of the estimated surface normal candidates in the existence of various noise sources 87
4.6 Examples of the surface reconstruction of a FFCS class . . . 95
4.7 Examples of the surface reconstruction of a RFCS class . . . 96
4.8 Examples of the surface reconstruction of an adult class . . . 97
4.9 Evaluation of surface reconstruction sensitivity to different depth 98
4.10 Comparison of the reconstructed surfaces with different depths 99
4.11 An example sequence of the 3D surface reconstructed from an adult class 100

5.1 Features from depth information 106
5.2 Spread axes information 107
5.3 The structure of a neuron 111
5.4 A typical example of a multi-layer neural network 112
5.5 The conceptual illustrations for two network topologies 113
5.6 Recurrent network topologies 114
5.7 A delay line with one tap 115

5.8 Classifier design 117

6.1 The proposed system implemented in a X-window based environment 120
6.2 Experimental setup 121
6.3 Typical samples of child seats 122
6.4 Various supplementary objects 123
6.5 Feature consistency 125
6.6 Classification error space with respect to different lengths of the tapped delay lines 128
6.7 Error space plot of the classifier with the weighted delay lines 128
6.8 Misclassification examples 131
6.9 Sample sequence for the surface distortion caused by severe motion 132

Abstract

This thesis introduces a novel framework for a real-time occupant detection system capable of extracting both *two-* and *three-dimensional* information using a *single imager* with *active illumination*. The primary objective of this thesis is to demonstrate the feasibility of such a *low-cost* classification system with comparable performance to multi-camera based stereo vision systems. Severe illumination conditions characterised by a frequent and wide illumination fluctuation are also challenging problems addressed in this work. The proposed system is designed to solve a problem of classifying three occupant classes being an *adult*, a *forward-facing child seat*, and a *rear-facing child seat*.

DoubleFlash is employed to eliminate the influence of ambient illumination and to compress the optical dynamic range of target scenes. The idea underlying this technique is to subtract images flashed by different illumination power levels. The extension of this active illumination technique leads to the development of a novel shadow removal technique, called *ShadowFlash*. By simulating an artificial infinite illuminating plane over the field of view, the technique produces a shadowless scene without losing image details by composing multiple images illuminated from different directions. The ShadowFlash technique is then extended to the temporal domain by employing the *sliding n-tuple strategy*, which is introduced to avoid the reduction of the original frame rate.

A modified *active contour model*, facilitated by morphological operations, extracts the boundary of the target object from the shadow-free scenes produced by the ShadowFlash. Based on the brightness information of the image *triplet* generated by the DoubleFlash, the orientations of the object surface at pixel points are estimated by the *photometric stereo method* and integrated into the 3D surface by means of global minimisation. The boundary information is used to specify the region of interest to reconstruct. Investigating both the two- and three-dimensional properties of vehicle occupants, 29 features are defined for the training of a neural network. The system is tested on a database of over 84,000 frames collected from a wide range of objects in various illumination conditions. A classification accuracy of 98.9% was achieved within the decision-time limit of *three* seconds.

Declaration

I hereby declare that this thesis has been composed by myself and that the research reported herein is my own, except where explicit reference has been made to the work of others. I especially wish to express my sincere thanks to Dr. Koch for his contribution in Chapter 2. Parts of this work have been published in collaboration with research fellows and colleagues in the following papers and patents:

- J.J. Yoon, C. Koch, and T.J. Ellis. Shadowflash: an approach for shadow removal in an active illumination environment. In *British Machine Vision Conference (BMVC02)*, pages 636–645, Cardiff, UK, August 2002. BMVA Press.
- J.J. Yoon, and T.J. Ellis. Real-time occupant detection system using active illumination. In *British Machine Vision Conference (BMVC04)*, pages 497–506, London, UK, September 2004. BMVA Press.
- J.J. Yoon, C. Koch, and T.J. Ellis. Vision Based Occupant Detection System by Monocular 3D Surface Reconstruction. In *IEEE Conference on Intelligent Transportation Systems (ITSC2004)*, pages 435–440, Washington, D.C., USA, October 2004.
- J.J. Yoon, C. Koch and L. Eisenmann. Verfahren und Vorrichtungen zur Schattenkompensation in digitalen Bildern[1]. National Patent DE 102.50.705, October 2002.
- J.J. Yoon, S. Weidhaas and M. Bauer. A device for the detection of an object on a vehicle seat. National Patent PA 040.43.418 DE, patent pending, September 2004.

[1] Methods and equipments of shadow compensation in digital images.

Acknowledgement

First and foremost, I owe my most sincere gratitude to my supervisor, Prof. *T. Ellis*, for his invaluable research and personal advice and for his direction throughout the preparation of this thesis. His understanding, encouraging and personal guidance also have provided a good basis for the present thesis.

My sincere gratitude is also due to my research fellow, Dr. *C. Koch* for our numerous spirited discussions, as well as serving as a sounding board to clarify concepts and ideas. His wide knowledge and logical way of thinking have provided a remarkable influence on this work.

I would like to express my deep and sincere gratitude to Mr. *N. Lii* for his constructive criticism and excellent advice during the preparation of this thesis. I could not have completed this thesis without his valuable comments and recommendations which elucidated many ambiguous points in the manuscript.

During this work, I have collaborated with many colleagues at BMW AG. I wish to extend my warmest thanks to all those who have helped me with this work, especially Mr. *U. Wagner*, Mr. *M. Bauer*, Mr. *S. Weidhaas*, Mr. *S. Akisoglu*, and Mr. *J. Mahalek*. The financial support of the BMW Group is gratefully acknowledged.

Finally, I owe my loving thanks to *my mother*. Without her encouragement and understanding, it would not have been possible for me to finish this work.

Munich, Germany, September 2004

Jason JeongSuk Yoon

The majority of this research was sponsored by BMW and performed at the BMW Group Research and Innovation Center in Munich, Germany. The opinions expressed herein are those of the author and do not necessarily represent those of BMW.

Chapter 1

Introduction

1.1 Aim of this work 2
1.2 Vision-based occupant detection systems 3
1.2.1 Motivation 3
1.2.2 Occupant detection systems 5
1.2.3 State of the art 7
1.3 System overview 11
1.4 Organisation of the thesis 13

1.1 Aim of this work

This thesis is concerned with the implementation of a classification system for occupant detection by recovering three-dimensional surfaces using a monocular imager with multiple illuminations in sequences of two dimensional images. As the demand for sophisticated vision-based applications increases, the reconstruction of three dimensional shape, also called *surface reconstruction* becomes more interesting for those systems due to its apparent advantage in feature selection.

Since the field of three dimensional imaging was introduced, there have been a number of difficulties for employing three-dimensional surface reconstruction into practical vision systems: the calculation of the surface generally involves a combination of time consuming tasks such as intensive searching, transform of domains, and/or triangulation. And in the early days of machine vision the computational cost necessary for the surface reconstruction easily overwhelmed the maximum power of available practical computing systems. However, as the performance of the microprocessor has been rapidly increased according to Moore's law over decades while the optimisation of algorithms has improved, real-time surface reconstruction based vision systems have become no longer imagination.

Although the vision and processor technology has overcome the limitation in terms of real-time functioning hardware implementation, there are still some cost sensitive applications which are not able to take an advantage of the advanced vision technology due to practical reasons. For example, vehicle cabin surveillance system could be one of the most cost-sensitive applications. To employ a binocular-based stereo system only for recovering the three-dimensional shape of a passenger may not be persuasive enough for conservative customers. Vehicle manufacturers are still seeking an alternative technology which could guarantee the comparable performance to those of systems using multiple imagers while the implementation costs are considerably less. The fact is especially true for applications where the generation of three-dimensional surface in the full resolution with the maximum frame rate is unnecessary and a *low frame rate* version with the *degraded resolution* is acceptable for the purpose.

Another issue that makes the realisation difficult is that practical vision systems are often exposed to an uncontrollable lighting environment. Most modern vision algorithms assume that the supreme performance is guaranteed only if the algorithm operates in a particular illumination condition. However, in reality, many vision applications must operate over wide and frequent illumination variations which increase the uncertainty of the individual image processing tasks.

For example, mainstream CCD based and most of the emerging CMOS based imaging sensors provide a high optical dynamic range of 48 to $60dB$ [51].

This dynamic range is obviously not sufficient for scenes involving extreme contrast (c.f. a scene with both bright and deep-shaded areas), so that the images obtained by those imagers often fail to capture the details of the scene. Another example could be cast shadows in active lighting conditions, which are usually assumed to have a negligible influence on the system performance. In fact, the cast shadows often degenerate the overall performance of vision systems by being misclassified as an imaginary object or artifactual parts of an object.

In order to address these aforementioned problems, this work proposes an alternative system capable of extracting both *two-* and *three*-dimensional information from the field of view in *real-time* using the minimal number of imaging sensors with some supplementary illuminations. The focus of this thesis is mainly to demonstrate the *feasibility* of such a system by employing machine vision techniques and additionally to find solutions for potential problems accompanied with the practical implementation in extreme illumination conditions.

The vision-based occupant detection system is chosen as a target system, since the design of this system involves most of the common problems experienced by vision based applications. However, most of investigations and experiments in this thesis are not limited to motor vehicle applications but also applicable to other machine vision systems in high dynamic range environments such as a industrial inspection, building surveillance, etc. Before launching into a detailed description of the proposed system, the history and fundamental concepts in the domain of occupant detection is briefly described in Section 1.2.

1.2 Vision-based occupant detection systems

1.2.1 Motivation

Airbags have saved several thousand lives worldwide so far and protected numberless passengers from serious injuries [75]. However, in late 1996, reports began to surface of airbags causing serious or fatal injuries in certain circumstances. During a frontal automobile collision, momentum can carry an unrestrained passenger forward until impact with the vehicle interior structure. A properly deployed airbag provides a much softer impact surface than a steering wheel or other interior surface, but the airbag must be fully inflated before impact to provide the maximum benefit. For the nominal 48 km/h barrier crash event, the time from impact to full deployment is generally in the order of 50 milliseconds. During this time, the occupant will move about 5 cm forward relative to the vehicle. It takes the airbag about 30 milliseconds to deploy, leaving 20 milliseconds for the sensor system

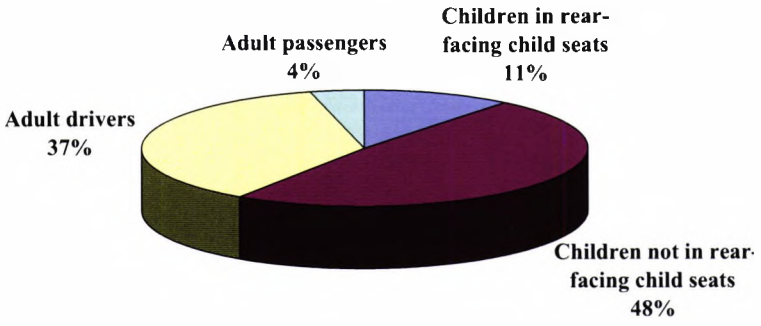


Figure 1.1: The statistics of the fatalities caused by airbag deployment from 1990 to 2000. It shows that 175 fatalities have been accounted including: 19 children in rear-facing child seats, 85 children *not* in real-facing child seats, 64 adult drivers and 7 adult passengers [100].

to determine the crash profile and begin the deployment procedure. Because of the short amount of time allowed for deployment for the nominal crash pulse, the airbag must inflate aggressively and an occupant who comes in contact prematurely with the airbag in the early stages of its deployment is at risk of injury from the airbag.

Figure 1.1 shows the statistics of the fatalities caused by the airbag deploy. Since 1990, NHTSA [1] has recorded 175 fatalities as a result of an airbag deployment by the end of 2000 in the U.S. 104 of these deaths have been children while the remaining 71 have been adults. The 86 children who died during airbag deployments were front seat passengers. The NHTSA has concluded that 76 of these children were totally unrestrained or improperly restrained, including ten who were only wearing their lap belts, effectively negating the advantages of a safety belt. Figure 1.2 shows a simulation of an unbelted child during an airbag deployment. Infants placed in the front seat of a car even in a rear-facing child seat have accounted for 19 deaths. Placing a child in the front seat of a car in a rear-facing child seat carries serious risks because the child's head is too close to the airbag compartment.

Furthermore, for the 18 people who were properly restrained out of 71 adult fatalities, NHTSA's investigations indicate that eight of these people were small stature females who were positioned close to the airbag housing. Two other fatalities involved men, who both lost consciousness before impact, thus moving their bodies closer to the airbag compartment. Most airbag systems assume that an occupant is a medium weight male (75%) in mid-seating position, whereas in reality 70% of the passengers are smaller and sitting closer to the airbag pod. Hence, adults can also be endangered by airbags if they take up an adverse seating position or attitude, called

[1] National Highway Transportation and Safety Administration

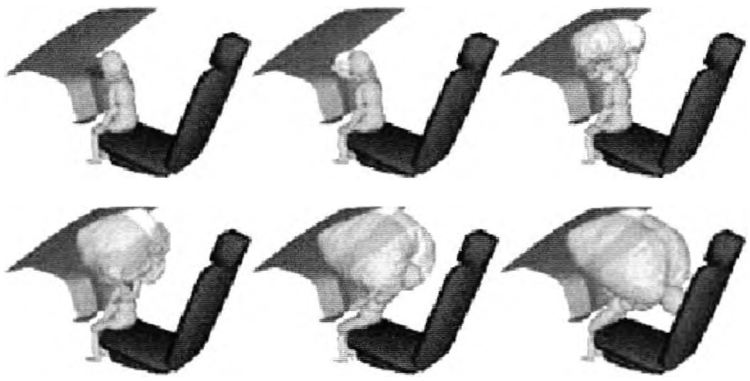


Figure 1.2: Simulation of an unbelted child during an airbag deployment.

Abbreviation	Occupant class description
FFCS	Forward-facing child seat
RFCS	Rear-facing child seat
PCSP	Person in correct seating position
POOP	Person out-of-position
NPOS	Noting present on the seat
ODFC	An object which does not fit to the other classes

Table 1.1: Description of the predefined occupant classes

out-of-position [100].

1.2.2 Occupant detection systems

The federal response was to propose new regulations in order to avoid passenger injuries due to restraint systems. Recently, the Federal Motor Vehicle Safety Standard (FMVSS) 208 set out by the NHTSA announces that nearly 100 percent of all automobiles sold in U.S. must have the ability to automatically control the deploying power of airbags based on crash type, crash severity, occupant type and size, as well as seat belt usage, starting with the 2006 model year [99]. Accordingly, almost all vehicle manufacturers and most automotive component suppliers are actively developing so-called *smart* or *advanced airbags* which should eliminate the risks produced by current airbag designs. It is likely that some form of these advanced airbags will be introduced into the market within the next few years.

The essence of this system is to recognise a passenger in the front passenger seat and adapt the airbag deployment with respect to the predefined

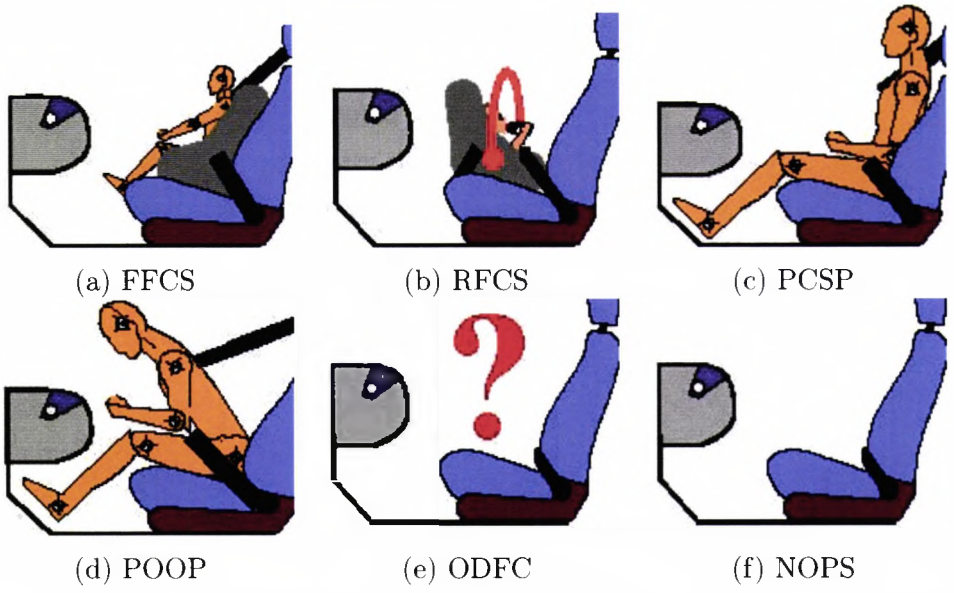


Figure 1.3: Illustrations for the six occupant classes: (a) FFCS, (b) RFCS, (c) PCSP, (d) POOP, (e) ODFC and (f) NOPS

occupant profiles based on their extensive statistical data. Table 1.1 introduces six occupant classes and their abbreviations which are most commonly used and illustration for each class is shown in Figure 1.3. To protect a correctly seated person (PCSP) the optimum protective airbag modus is to inflate an airbag to full size within a minimum of time, while only a reduced amount of power is allowed to blow up the airbag for both child seats (FFCS and RFCS) and persons in out-of-position (POOP). This can be realised by employing the recently introduced *multiple stage* or *de-powered* airbag systems which have multiple ignitors providing different inflation power according to the occupant profiles. If the system detects either an undefined object (ODFC) or the absence of an occupant (NOPS), it prevents airbag deployment to avoid the high cost of replacement.

The overall system is often explained as the combination of two subsystems being *occupant detection* (OD) and *out-of-position detection* (OOP or OOPD) systems. In order to manage situation-appropriate airbag deployment, an occupant detection system automatically detects the presence of an occupant in the passenger seat and categorises it into the predetermined occupant classes while an out-of-position system constantly monitors the occupant and initialises the airbag deployment procedure according to its behaviour if the occupant has been classified as a *person*.

1.2.3 State of the art

Due to the high safety relevance of this application the system must be designed to be predictable, error tolerant and accurate. Many different approaches on established sensor technology have been tested to realise such an advanced airbag system in recent years.

Manual switches

The most intuitive and cheapest solution to deactivate the passenger-side airbag is to implement a simple on/off switch being toggled manually by the driver or the passenger itself. Although it is not an automatic occupant detection system, it still helps to suppress unwanted airbag deployment especially in case that rear-facing child seats are mounted on the front passenger seat. The major disadvantage is that the decision is completely dependent on the driver so that he or she may involuntarily forget to switch off the airbag deployment when they drive with their child and it could put the child in great danger in the event of a collision. Another shortcoming of this approach is the fact that it is not possible to actively cope with the time-varying situations such as a passenger in out-of-position.

Load sensor systems

Most advanced airbag systems currently in service are usually based on a force sensitive sensor matrix in the seat. If a force is applied on the sensor matrix the resistance of each sensor changes, and the occupant sit-in pressure profile is measured and used for classification. By measuring and reacting not only to sit-in weight, but also the buttock print and positioning, centre of gravity distribution, variation of the seat assembly and components, and tolerances due to temperature and humidity, the system offers relatively high performance especially in cooperation with the seat belt tension detecting sensors.

The major drawback of these systems is that each new seat construction or modification implies an expensive re-design and calibration of the sensor matrix. Movement caused by an *Active Seat*, which is an integrated passenger massage system by alternating fluid pads mounted within the seat, may also disturb the system in making a right decision.

Transponder systems

Alternative method to provide automatic child seat recognition capability is to implement a simple transponder consisting of a transmitter and receiver coil in combination with an electronic control units into a child seat so that the child seat can be always appropriately detected by interacting with

the passenger seat transponder. More importantly, if two transponders are integrated, it could be possible to detect the orientation of the child seat. However, this approach would introduce a price increase to the child seat in a cost-sensitive industry. Availability in only high-end child seats on the other hand, would not meet government requirements. Also the detection of the out-of-positioned person still remains unresolved.

Machine vision based systems

As manufacturers began to develop various *occupant detection systems*, the vision techniques have attracted much attention due to their superior adaptability to various vehicle cabin environments as compared to the other mechatronic methods, and a number of machine vision based approaches have been studied to resolve the airbag suppression decision problem in recent years [55, 73, 57, 62, 72, 101, 23, 51]. The following section gives a brief overview of already published approaches for monitoring the interior of vehicles for reasons of safety and convenience.

Single camera approaches Park proposed an optical occupant detection based on the idea of searching for a human face in the scene [73]. If an adult face can be detected properly in case of a crash the airbag should be inflated, otherwise it should not. A single monochrome camera is employed to capture image sequences and a set of eigenfaces is created by using the principal component analysis (PCA) for face-image matching. The idea of this approach can be easily shared with the other important vehicle cabin applications such as a *driver drowsiness detection*. Nevertheless, it is limited in its real application due to the fact that a straight look into the occupant's face is not guaranteed all the time and the large size and orientation variations of the passenger's face make it difficult to correlate them with the normalised eigenfaces. High illumination fluctuations present in a vehicle are another factor the system must overcome to be practical. Furthermore, a more accurate distinction between FFCS, RFCS and POOP is necessary for the optimised control of multi-stage airbag deployment.

Farmer introduced a low cost, high reliability occupant classification system using a single grey-scale camera in [23]. A four-class problem with the classes being rear-facing infant seat, child, adult and empty seat was addressed based on a database of over 21,000 real-world images, collected over a period of 4 months in order to prevent the system from being adapted to a particular illumination condition. Using supplementary infrared illumination, a CMOS-based camera with a wide-field-of-view lens captured a single image and shape features were extracted from the region of interest provided by a preceding segmentation task. A set of multiple k-nearest neighbour classifiers trained using over 150 features, resulted in an overall classification accuracy of better than 95%.

Another grey-scaled monocular camera based occupant detection system is proposed by Koch in [51]. Combining a set of images flashed with different radiant intensities, the dynamic range of the scene was successfully compressed and the illumination offset produced by the ambient light was also completely eliminated. After the adaptive thresholding applied on the similarity comparison result between the input and reference image, Fuzzy logic is performed on the features reflecting the geometrical properties of an occupant and the overall classification rate reached at 94%.

Stereo vision using multiple cameras A stereo vision based occupant detection system for the airbag deployment is proposed by Krumm in [55]. Two different experiments were performed using a single camera and a binocular stereo camera in order to assess the advantage of having range data from stereo images for the classification over intensity images. The experiment proved that the binocular stereo technique is less sensitive to varying illumination conditions and the range data can be used for giving important clues to estimate the position of the occupant. The prototype images used for image matching was trained with a set of test images of empty seats and seats occupied by rear facing child seats using principal components analysis, and the classifier compared input images with the prototypes by matching their eigenvectors. Therefore the system was only able to distinguish a limited set of RFCS, an empty seat or an object which does not fit to the RFCS class, although the classification rate reached at 95.1% on a test of 890 images.

In [101] Trivedi presented a stereo and *thermal infrared* (TIR) video based real-time vision system for both occupant detection and out-of-position detection. A comparison was made on a frame-by-frame bases between range data captured by a *trinocular stereo camera* and the TIR images, in detecting the presence of an occupant as well as tracking the head location in the background-removed disparity data. The experimental results showed that the TIR system (93%) had relatively higher performance than the stereo approach (86%), although the TIR camera exhibited some undesired characteristics such as the change of the intensity mapping from skin temperature over time. Despite the success of this test, further testing was imperative since the comparison test of those system included only one subject at a particular time of day in particular weather. Furthermore, thorough cost analysis would be imperative to the economies of introducing a TIR sensor.

Alternatives An interesting study employing *structured lighting* was made by Lequellec in [57]. A system based on a CCD camera combined with a light beam matrix is developed to output the 3D surface shape of vehicle cockpit occupancy. An initial set of possible spot/beam matchings is deduced from epipolar constraints provided by the prior calibration of the relative camera/projector position. Then, using the topological constraints in the 2D

mesh of illuminated dots, a constraint propagation process eliminates most of the combinations possibilities from the initial set of matchings. Finally, the 3D corresponding points are then computed via triangulation from this matchings set. However, in reality, it is difficult to implement an accurate pattern using an infrared light source due to the constant vibration in the vehicle environment. Furthermore, such patterns may not provide enough resolution for object classification.

In [50] Klomark presented a number of potential machine vision based approaches for occupant detection systems and evaluated the usefulness of each technique. It is concluded that any simple image matching approaches, especially which are based on edge properties of an object, were not satisfactory. The experimental results showed that the robustness to illumination varying environments was essential for realising a practical occupant detection system and a combination of techniques should give enough reliability for safety applications.

Summary

These previous studies for vision based occupant detection can be classified into two categories depending on the number of cameras used in the system. In the earlier versions of occupant detection systems, single camera approaches were in demand due to the high cost of imaging sensors. However, such monocular systems did not provide sufficient 3D information necessary for functions such as the out-of-position detection, which is a supplementary task guaranteeing low risk deployment according to the position/pose of the passenger. As a consequence, the majority of occupant detection systems became more dependent on stereo vision techniques using multiple cameras.

Faced with the increasing demand for various vision-based in-vehicle applications, the growing number of cameras employed has come under serious scrutiny. For this reason, this research focused on developing a single camera system able to generate additional 3D information by using minimal supplementary active illuminations, in order to circumvent the higher costs of components and the complication of installation, maintenance and calibration. The primary objective of this thesis is to propose a novel framework mainly for, though not restricted to, the occupant detection system, as well as to demonstrate the feasibility of alternative systems with comparable performance to multi-camera based vision systems. These efforts resulted in the development of, (1) a framework capable of extracting three-dimensional information of an object with minimal hardware costs, (2) a few useful techniques to stabilise the illumination conditions, (3) a set of robust and efficient feature descriptions which characterise the size and position of the occupant, and (4) a simple pattern recognition module to classify the visual cues in categories, which can trigger the safe deployment logic of the

airbag system.

1.3 System overview

The proposed occupant detection system is designed to classify an object in a vehicle for facilitating the airbag control module. The airbag responses of two occupant classes including ODFC and NOPS were assumed that they do not necessarily have to be detected separately if considering only the safety requirement. Assuming PCSP and POOP classes were supposed to be distinguished in a *out-of-position detection system* by sharing the classification results made in occupant detection, the target occupant classes in this thesis were reduced to three classes being a *front facing child seat*, a *rear facing child seat* and an *adult*. Figure 1.4(b) illustrates the state transition between the occupant detection and out-of-position modules. The out-of-position detection is activated only if the object is classified as an *adult*, which is the *merged* class continuously observed after the classification in order to detect a person in out-of-position.

Figure 1.4(a) shows a basic framework of the system. A 12-bit high dynamic range monochrome imaging sensor with the resolution of 324×244 at 30 Hz was employed for the image capture. *Three* infrared light sources triggered by a 2-bit gray code signal were used, flashing in sequential order. The gray code signal is also synchronised with the trigger clock in the imaging sensor so that each frame must be captured under the pre-assigned illumination conditions. Accordingly, *four* types of images having different illumination are consecutively obtained during the acquisition: three images by each light source (plus ambient illumination) and one with only ambient illumination.

After eliminating the lens distortions, the image sequence is delivered to the *DoubleFlash* module originally introduced in [52], which eliminates the ambient illumination fluctuations by subtracting two images exposed by different illumination powers. Being facilitated by a three-stage delay buffer, the DoubleFlash method completely removes the ambient illumination and produces three images per clock cycle. These images are used by *ShadowFlash* [116] to compose a shadow-free image by simulating a virtual light source having an infinite extent. A *deformable contour model* is then applied to this shadow-removed image in order to extract the boundary information of the object. By fusing the three images created by the DoubleFlash, the *photometric stereo method* reconstructs the 3D surface of the object with the help of the segmentation result. Finally, a 29-dimensional feature vector defined using both 2D and 3D information is utilised to train a *neural network* to make a single decision per each frame. *Tapped delay lines* are introduced to filter noise in both the input and output of the network in order to increase temporal consistency of the classification results.

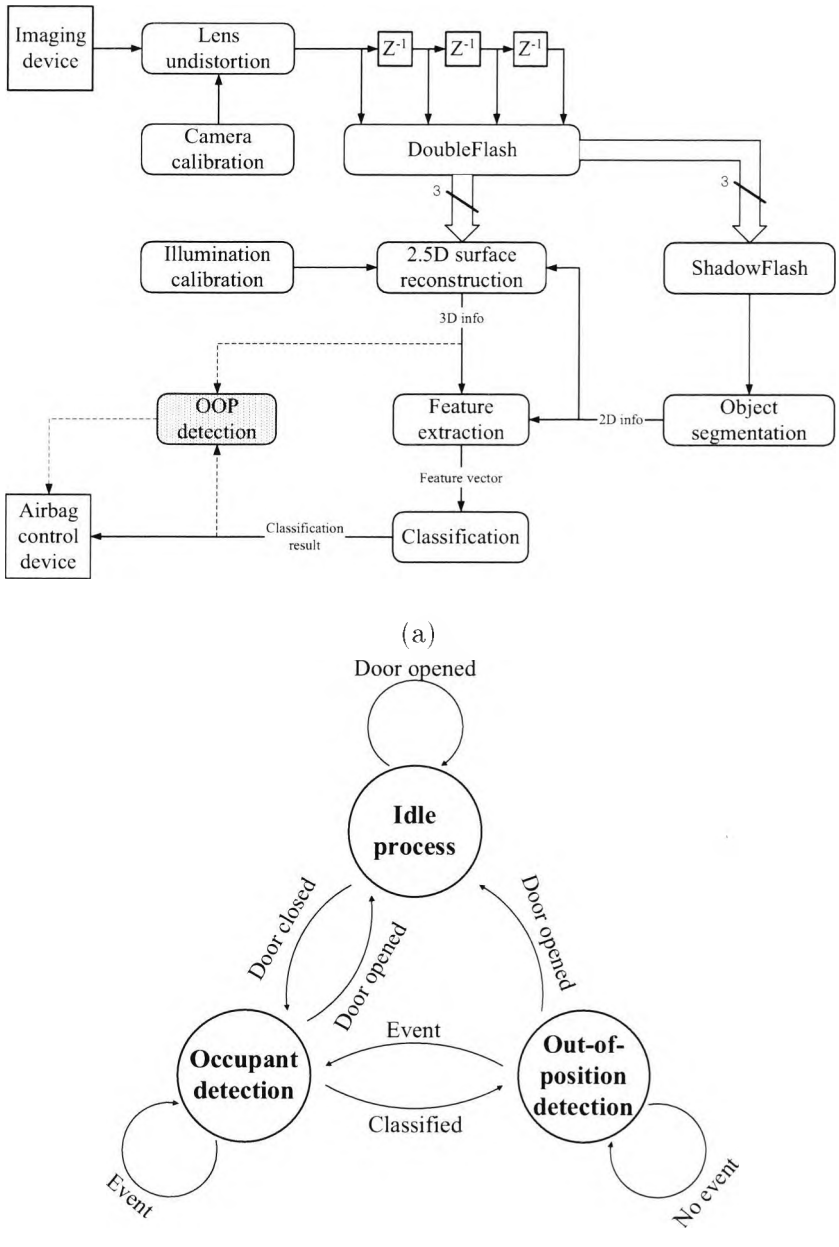


Figure 1.4: System overview: (a) the structure of the proposed system in conjunction with the out-of-position detection system, and (b) the state transition diagram of the overall system. The transition 'Event' occurs when any dramatic change happens in the field of view, such as any abrupt change of classes.

1.4 Organisation of the thesis

In Chapter 2, the problem of obtaining illumination-stabilised images in high dynamic range environment is discussed. Solutions for the improvement of robustness to the ambient illumination fluctuations and the minimisation of the effects caused by the cast shadows are introduced. Extracting two-dimensional information using *active contour models* is presented in Chapter 3 while Chapter 4 discusses the three-dimensional surface reconstruction method with real-time video sequences based on the *photometric stereo method*. In Chapter 5, a novel feature set collected from both the two- and three-dimensional information as well as the design strategies of the classifier are introduced. Experimental results considering the occupant detection system will be shown, followed by a discussion about the expected problems in perspective of the practical system realisation in Chapter 6. Finally, a summary and the conclusion of this thesis is presented in Chapter 7.

Chapter 2

Acquisition and pre-processing

2.1 Motivation 16
2.2 Sensors and illuminations 17
2.2.1 Optical dynamic range 17
2.2.2 Imaging sensors 18
2.2.3 Active illumination 24
2.3 Image enhancement. DoubleFlash 24
2.3.1 Introduction 24
2.3.2 Offset reduction 26
2.3.3 Dynamic range compression 27
2.3.4 Experimental results 28
2.4 Shadow removal: ShadowFlash 28
2.4.1 Motivation 28
2.4.2 Analysis 30
2.4.3 Shadow removal 31
2.4.4 Experimental results 38
2.4.5 Discussion 41
2.5 Precis 42

2.1 Motivation

Although innumerable efforts have been made to employ the sophisticated machine vision algorithms to the industrial applications, less attention has been directed to the image sensing techniques. Since the picture quality completely relies upon the optical sensor characteristics, the design and/or selection of an appropriate imager is one of the most essential procedures in the overall system design in order to guarantee the optimal performance for the individual image processing tasks. Especially for some vision applications which operate in an unrestricted illumination conditions experiencing extreme contrasts and frequent change of light conditions both spatially and temporally, the system performance will not be satisfactory unless the intensive investigation for seeking a suitable imaging sensor is considered from the beginning of development.

Pre-processing operations, sometimes referred to as image restoration and rectification, are intended to correct for sensor-specific radiometric and geometric distortions of data. Radiometric corrections may be necessary due to variations in scene illumination and viewing geometry, atmospheric conditions, and sensor noise and response. Each of these varies depending on the specific sensor and platform used to acquire the data and the conditions during data acquisition. Furthermore, it may be desirable to convert and/or calibrate the data to known (absolute) radiation or reflectance units to facilitate comparison between data. Variations in illumination and viewing geometry between images for optical sensors can be corrected by modelling the geometric relationship between the object of interest, the light sources and the sensor. It is also often required to mosaic multiple images from a single sensor while maintaining uniform illumination conditions in order to be able to more readily compare images collected under different lighting conditions.

This chapter begins with a discussion on the available imaging sensor technology in conjunction with illumination issues at present and in the near future with regard to their usability for high dynamic range environments. It continues with the introduction of novel techniques for solving the problems which often occur in vision based applications, especially operating in high dynamic range environments: The *DoubleFlash* approach [52] is used for capturing a scene under adverse lighting conditions without losing image details by employing active illumination. The novel *ShadowFlash* technique that produces a shadow-free scene by approximating an artificial light plane with an infinite extent using multiple spot-light sources is also introduced.

Graphics: C.Koch [51]

(a) (b)

Figure 2.1: Typical examples of image detail lost due to the limited dynamic range of an imager: (a) a sample image overwhelmed by the ambient illumination, and (b) a scene captured by an imager with sufficient optical dynamic range.

2.2 Sensors and illuminations

2.2.1 Optical dynamic range

As machine vision tries to leave the ideal conditions in laboratory environments with controlled illumination situations, the importance of the *optical dynamic range* for image processing has come into focus in recent years. Optical dynamic range could be defined as the range of irradiance amplitudes over which an imager can acquire the scene without unacceptable distortion of scene details. The optical dynamic range is usually expressed in dB and calculated as follows:

$$\Delta_{dynamic\_range} = 20 \cdot \log\left(\frac{E_{max}}{E_{min}}\right) \tag{2.1}$$

where E_{max} and E_{min} are the maximum and minimum irradiance, respectively.

Depending on the definition of the domain in which the irradiance measurements are made, the optical dynamic range can be further split into either *global* or *local* dynamic range. The global dynamic range means the amount of the light fluctuations for a certain amount of *time*, while the local dynamic range simply stands for the irradiance range *within a scene* at a specific time. If the dynamic range of an imager does not fit to the local dynamic range of a scene then the details of the scene shown in the image will be degraded. An example of the scene where the dynamic range of a imager is overwhelmed by the ambient illumination is shown in Figure 2.1.

Considering the limited dynamic range of the conventional CCD imagers, there are various situations where these camera systems cannot provide satisfactory image quality. Figure 2.2 shows a graph plotting illumination changes as a function of time. The irradiance power inside of a vehicle was measured during motorway drive in the late evening in Germany. Several signal peaks occurred due to the reflections caused by incident rays within the interior while the signal level went down by a factor of several hundreds after sunset. Another irradiance measurement shown in Figure 2.3 was made at a parking place by night. The maximum irradiance signals were caused by either active interior lighting or the headlights of a closely approaching car. The driving route of the last example graph shown in Figure 2.4 includes two tunnels where extreme changes of irradiance were detected as the car went through the tunnels.

An interesting experiment assessing the dynamic range of vehicle interior was made by Koch [51]. After the intensive measurements of irradiance powers performed in various time and places, it was shown that the maximum global dynamic range for the interior of a vehicle could easily reach 191dB and it is definitely beyond the abilities of present imagers. Consequently, it is necessary to find out the methods with which an image without losing scene details can be captured in an high dynamic range environment.

2.2.2 Imaging sensors

CCD- vs. CMOS-based imagers

Most modern electronic cameras use the CCD[1] technology, although MOS[2]-based imaging sensors are based on technology developed earlier in the 1960's. The main reason for the failure of the early MOS-based imagers was due to the fact that the size of the MOS pixel was very large so that only limited number of imaging cells were able to be placed in a chip. In contrast, the advantage of CCD technology was able to create much smaller pixels with the same structure size than the MOS-based imagers. Considering that the size of each imaging cell determines the signal-to-noise ratio of each pixel, it was obvious that the picture quality of CCD-based imagers was superior compared to the MOS imagers at the time. However, as the CMOS[3] integration technology has been replacing the MOSs for last decades, the size of the individual pixel has been decreased to dimensions comparable to those of CCD imagers. Accordingly, the image quality of

[1] Charge Coupled Devices

[2] Metal-Oxide Semiconductor

[3] Complementary Metal-Oxide Semiconductor logic uses a combination of p-type and n-type metal-oxide semiconductor field effect transistors (MOSFETs) to implement logic gates. An advanced version of the MOS technology.

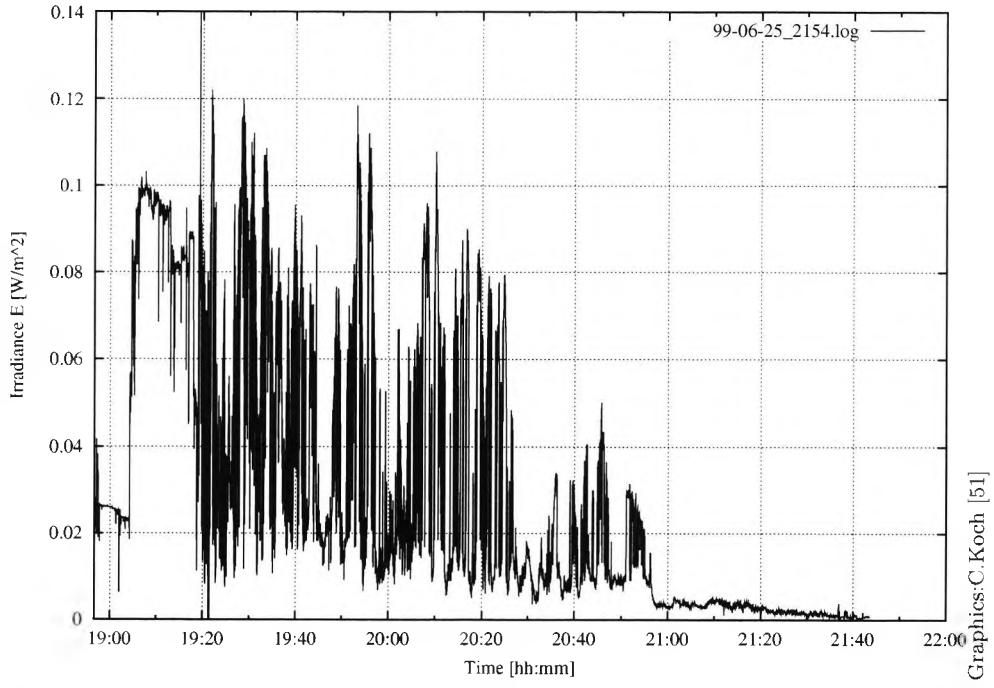


Figure 2.2: Irradiance variations during motorway drive. The irradiance level goes down as the sun sets [51].

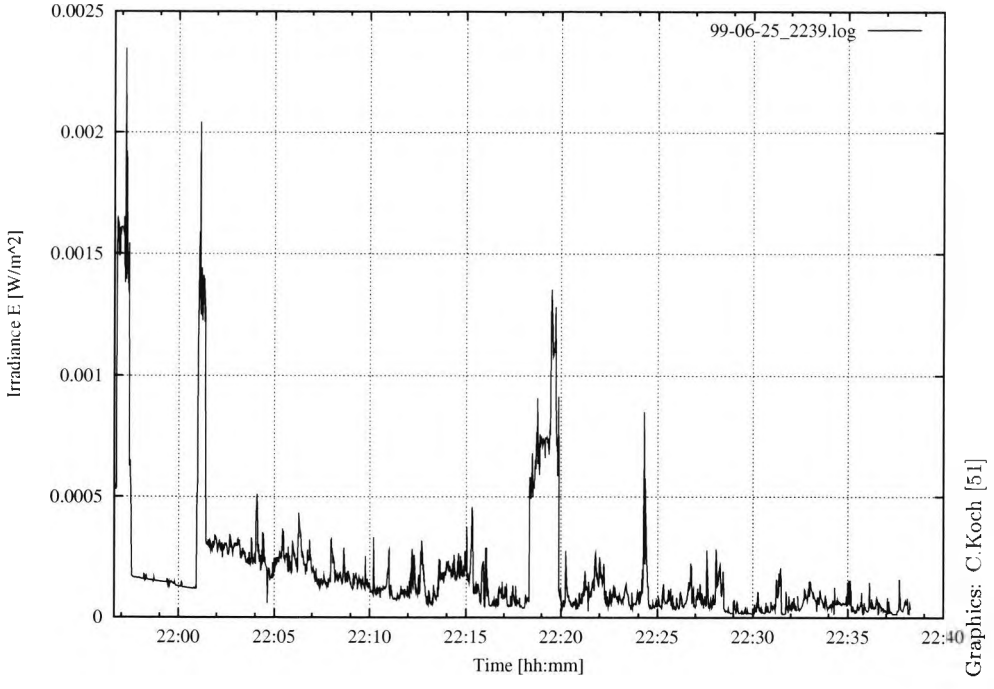


Figure 2.3: Irradiance variations in a parking lot. Sudden change of irradiance is possible due to the external illumination sources such as the headlights from the other vehicles [51].

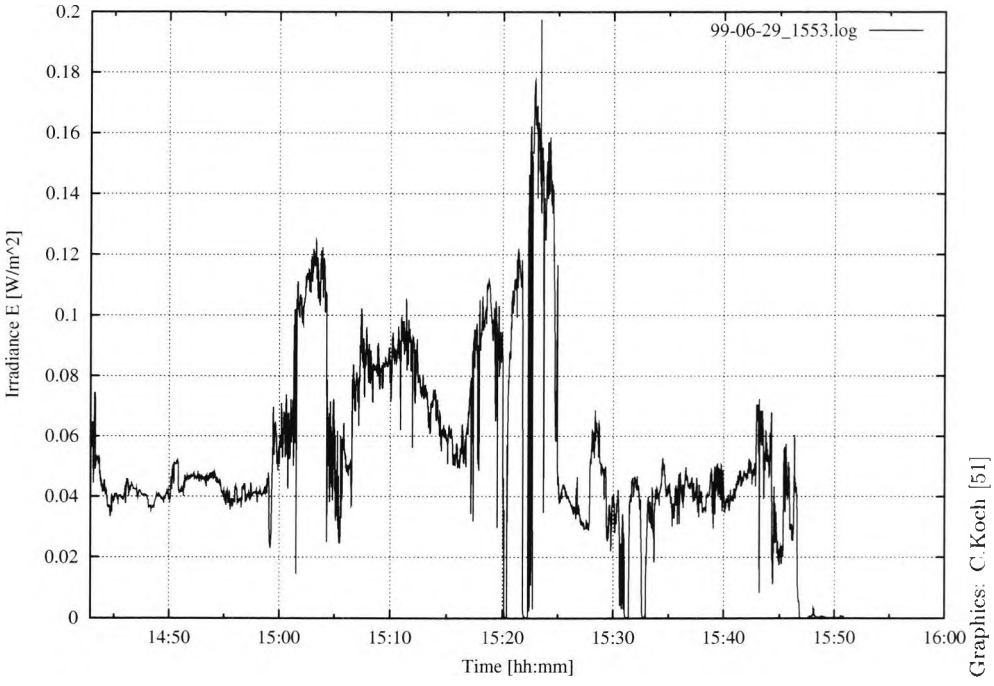


Figure 2.4: Drive through the city of Munich including tunnels [51].

CMOS sensors also has been dramatically improved, almost as good as the mainstream CCD sensors nowadays.

Advantages of CMOS sensors

CMOS sensors have a couple of significant advantages against the CCD sensors. Detailed comparisons between available sensor technologies were made by Koch in his thesis [51] and the following paragraphs present come of the major benefits for using CMOS-based sensors.

Active pixel sensors [4] Although the complex pixel structure of CMOS cells results in the degradation of image quality, it also could be considered as the major advantage of CMOS based imagers when employing an sensing-adaptability to high dynamic range (HDR) environments [51]. The extended high dynamic range can be achieved by either employing non-linear imaging elements or assembling a HDR image from a set of frames of a linear imager with multiple integration times, and those two ways can be only realised by the *active pixel sensor* (APS) properties of CMOS sensors [94].

[4] The photodetector and read-out amplifiers are implemented in each pixel, so that the sensitivity of each cell can be controlled individually.

Low power consumption Due to the requirement of only a simple power supply voltage for the CMOS imagers, typical CMOS based sensors consume only one fourth of the power of equivalent CCD cameras at the sensor level for an equivalent pixel clock [25]. Accordingly, they provide a much larger operating temperature range. The significantly reduced power consumption rate is important for certain stand-alone applications which have limited power resources for the operation such as mobile devices, automotive applications, etc.

System integrability The system-on-chip integrability with several supplementary signal processing functions is another important advantage of CMOS sensors. Due to the compatibility with standard CMOS technology it is possible to integrate all the necessary machine vision sub-routines into one small camera chip, and obviously this is not feasible to the CCD based sensors. The advancement of the CMOS manufacturing technology due to the growth of various CMOS-based application markets is another important factor that makes the price of CMOS imagers cheaper. These are the indispensable feature for realising low-cost vision-based systems.

Camera response characteristics

Presently available HDR cameras are usually based on CMOS technology. There are basically three different ways to realise a CMOS image sensor for HDR environments by extending its limited dynamic range, and the different photon-voltage transfer functions in linear, logarithmic and piecewise linear operation modes are shown in Figure 2.5.

Non-linear response The straightforward way to provide a high dynamic range is to compress the scene intensities by implementing a *logarithmic* voltage-current response of CMOS transistors. With this logarithmic intensity compression scheme, a CMOS camera is able to convert up to 120dB of intensities into a voltage range of a few hundred millivolts. However, this non-linear mapping results in distortion which is not recoverable by decompressing the intensity information since the textural information at low intensity level is already discarded from the acquisition due to the increased sensor noise level.

Linear response Due to the limited photo-sensitivity of CMOS sensors, it is not possible to obtain a high dynamic range directly from CMOS sensors with a linear response. The dynamic range for a linear pixel signal in an CMOS based imaging sensor is limited to less than 80dB due to the ASIC noise floor [51]. However, a dynamic range of over 100dB could be obtained by employing an approach to combine a HDR image using a set of images, of which each image is exposed for a different amount of time so that the different portion of bright-

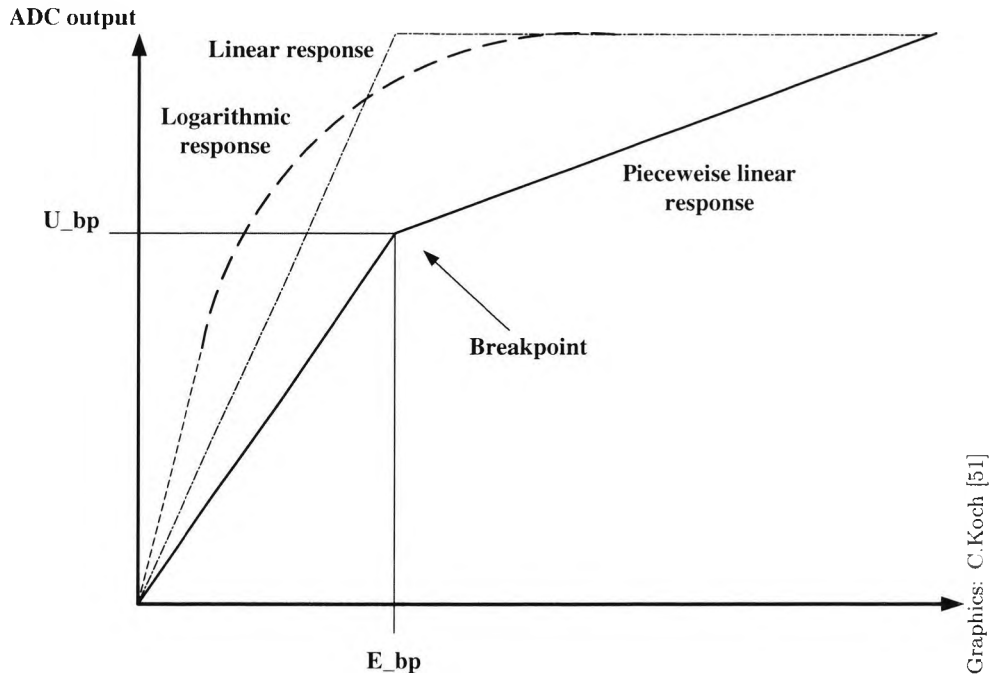


Figure 2.5: Comparison between different response functions of CMOS cameras.

ness information is assigned to each image. The images are assumed to be successively captured by a stationary camera with fixed focal length and linear response and finally merged into one image which has a greater dynamic range than a single snap image by an external logic [61, 41, 82]. Though the approach assumes the scene differences during image acquisition are negligible, there is still a possibility of having incorrect scene details caused by motion or ambient illumination fluctuations. And the frame rate is another factor which affects the quality of the composed HDR image.

Piecewise linear response A solution to reduce the complexity of composing a HDR image captured with a linear sensor while achieving sufficient image contrast at low light conditions is to employ a *piecewise linear* response [68, 108, 51]. In this technique the sensor's photosensitivity curve in each individual pixel is initialised at the beginning of acquisition. As the sensor begins to integrate photons, the photons are accumulated as in normal linear operation mode until a pre-determined voltage level is reached. If the level is not reached this may continue over the full integration time. If the signal level exceeds the pre-determined threshold the effective integration time is reduced, providing lower sensitivity for the bright light level. By this scheme, higher contrast level and less noise at the low light level compared to

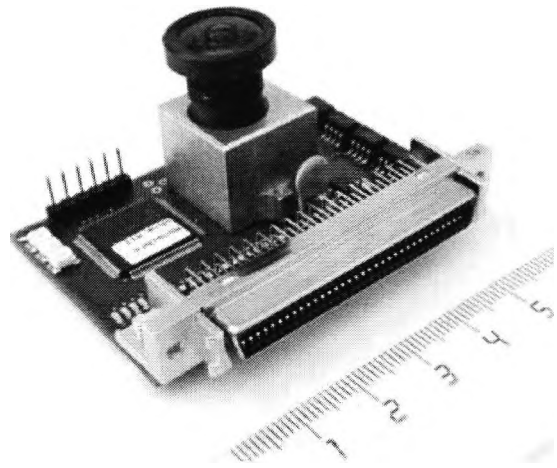


Figure 2.6: SollyCam version 3.0

the logarithmic pixel compression are guaranteed.

SollyCam

Since one of the key features in this work is to design a low-cost vision system and evaluate it in unconstrained illumination conditions, there was a need for designing a suitable camera system for high dynamic range applications such as a vehicle interior analysis introduced in Section 1.2. Concerning the unsatisfactory features of CCD-based imagers and the necessity of operating in NIR range, a customised CMOS-based camera called *SollyCam* was designed and used through most of experiments performed in this thesis. By providing increased flexibility to the latest version of the SollyCam, the usage of the camera is not limited to the occupant detection system but is also applicable to another outdoor machine vision applications such as various surveillance systems. Facilitated by modern CMOS technology, the sensor includes the completely integrated functions of timing unit, A/D converters, interface drivers, fixed pattern noise suppression logic, etc. Figure 2.6 shows the design of the imager used in this work, while the following list enumerates some key features of the SollyCam based on the CMOS sensor LM9618 from National Semiconductor [68].

- VGA resolution (648×488 pixels)
- High dynamic range scene over 100dB could be captured using its *piecewise linear response* capability.
- 12-bit grey-scale digital image output plus 4 additional reserve bits using RS422 interface which is programable by PLD configuration.
- Low power consumption rate (sensor/camera=160mW/2W), programable power-saving mode.

- Variable timing and snapshot mode including programable line, row and frame delays via I^2C interface.
- Pixel clock up to 12MHz (which provides 30Hz frame rate with the resolution of 324×244.)
- Guarantees 70% sensitivity within the near infrared (NIR) range at $\lambda \simeq 800nm$.
- Low-cost due to mass production.
- Operating temperature: -40 to 85°C
- Embedded trigger logics for controlling active illuminations.

2.2.3 Active illumination

It is most preferable to develop all the vision techniques independent from any illumination changes. However, many of modern machine vision algorithms are in fact based on the hypothesis of no illumination variations which often makes vision-based systems difficult to be adapted to real world situations. An image could be relieved from being disturbed by ambient illuminations with support of *active illumination*[5] and many non-lab vision applications experiencing unrestricted light conditions could overcome the illumination problems this way.

For applications where the active illumination must not be recognised by users, such as a vehicle interior monitoring task, *near infrared* (NIR) is often considered a suitable supplementary illumination due to its complete invisibility to human eyes as well as the feasibility of manufacturing low-cost light emitters. Using a proper bandpass filter with a centre wavelength of 778.4nm, all irradiation out of the passband can be completely blocked. In addition most of cameras based on the silicon technology are still sensitive at these wavelengths and the spectral power density of the sunlight decreases significantly beyond the NIR region. i.e. no special imager is required to get NIR images if an appropriate bandpass filter is employed. Figure 2.7 shows the spectral irradiance of the sunlight observed from the earth through the atmosphere while the typical transfer function of the NIR bandpass filter is shown in Figure 2.8.

2.3 Image enhancement: DoubleFlash

2.3.1 Introduction

Due to the limited optical dynamic range of conventional imagers together with the wide range of illumination fluctuations in realistic scenes, an illu-

[5]Illuminating a scene with controllable light sources is called *active illumination* while illumination influencing the scene which can not be manipulated is called *passive illumination*.

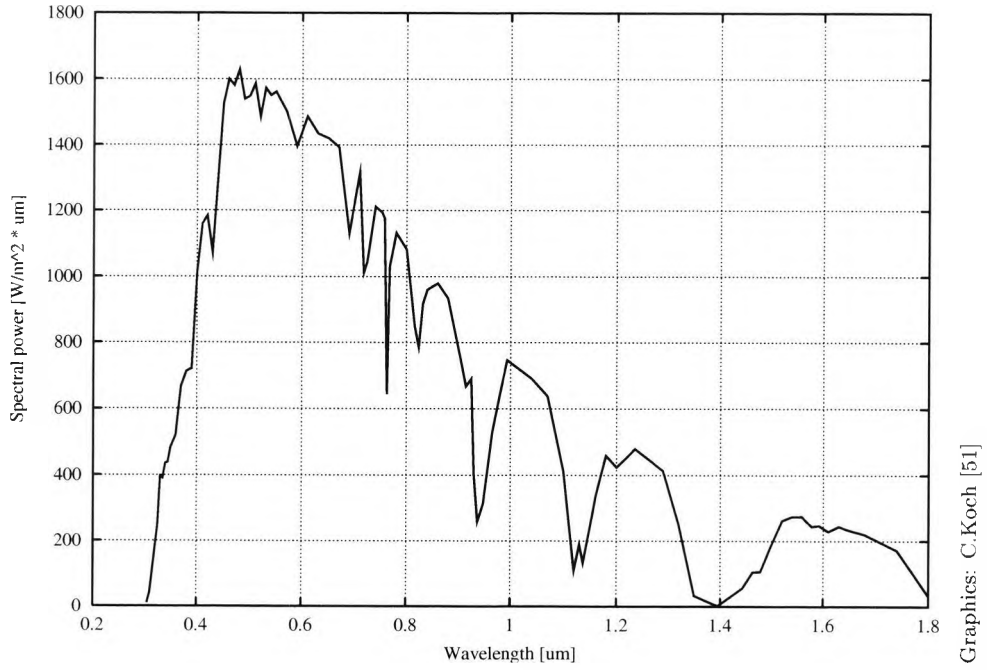


Figure 2.7: Spectral irradiance of the sunlight after atmosphere.

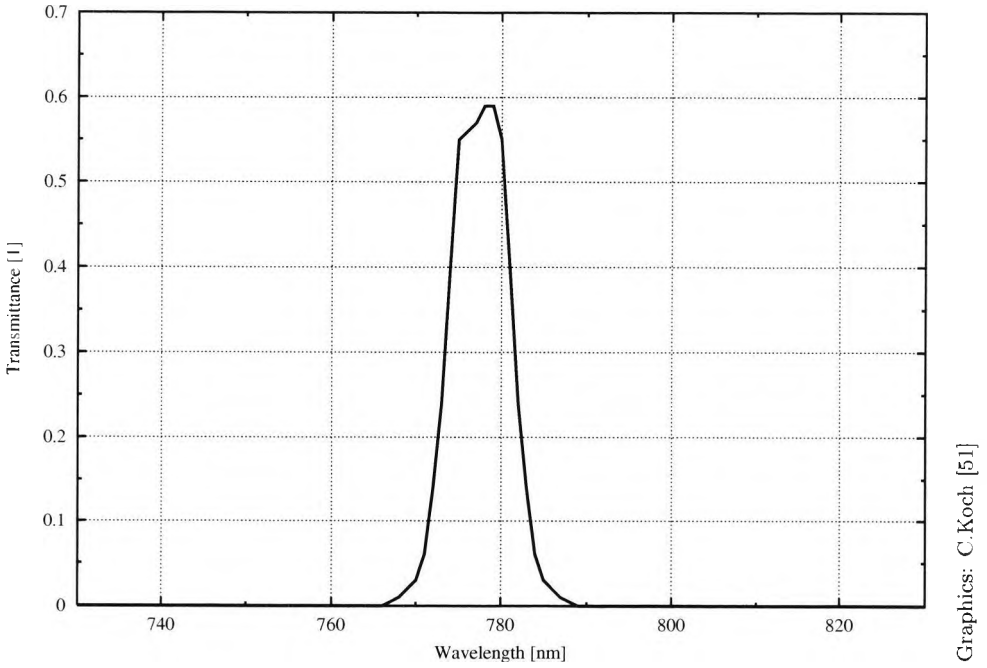


Figure 2.8: Transfer function of a typical NIR bandpass filter of which the centre wavelength λ=778.4nm and the bandwidth $\Delta\lambda$=10.7nm.

mination regulation technique called *DoubleFlash* for minimising potential illumination fluctuations by employing active illuminations is introduced in [52]. Additionally, this method has another effect that the optical dynamic range within the image is compressed compared to the image without supplementary illuminations. The basic idea of the method is to analyse a pair of image frames which are illuminated by light sources of different intensities. Finally, this technique is extended for the shadow removal approach discussed in Section 2.4.

2.3.2 Offset reduction

A sequence captured with varying illumination offset and noise usually requires to be refined with support of additional pre-processing steps for image analysis [53, 52, 74, 51], and sometimes these pre-processing tasks become one of the undesirable factors which hinder the efficiency of a system.

In order to achieve an offset and noise reduction and simplify the pre-processing procedures the DoubleFlash method employs two light sources having *different radiance intensities* E_{high} and E_{low}. Suppose that a digital image $I(n)$ only consists of the surface reflectivity $\rho_s(n)$ and irradiance power $E(n)$, the image $I_x(n)$ which is illuminated by a light source with irradiance E_x is defined as

$$I_x(n) = \rho_s(n) \cdot (E_x(n) + E_{amb}(n)) \tag{2.2}$$

where E_{amb} represents the ambient illumination irradiance.

An image I_{low} with only one illumination having irradiance power of E_{low} is acquired at time n, while the second image I_{high} with the illumination power of $(E_{high} + E_{amb})$ from the other supplementary light source at time $(n + \varepsilon)$ is sequentially captured. Assuming that the position of the camera is stationary over time, ρ_s would be constant for all frames. In this case the intensity levels of the captured images become only a function of irradiance E. As the time difference ε tends to zero, the subtraction between the images I_{high} and I_{low} yields an image which reflects only the difference between the received radiant powers of two input images while the influence of the ambient illumination E_{amb} is cancelled out. i.e. the output image I_{flash} becomes independent from the light fluctuations as shown in Equation 2.3.

$$\begin{aligned} I_{flash}(n) &= |I_{high}(n+\varepsilon) - I_{low}(n)| \\ &= \rho_{s,const.} \cdot |E_{high}(n) + E_{amb}(n) - (E_{low}(n) + E_{amb}(n))| \\ &= \rho_{s,const.} \cdot |E_{high}(n) - E_{low}(n)| \qquad (2.3) \\ &\quad qed. \end{aligned}$$

2.3.3 Dynamic range compression

The utilisation of supplementary illuminations compresses the local dynamic range of the flashed image. The dynamic range of an imaging sensor $\Delta_{dynamic\_range}$ is commonly defined as the ratio of its largest non-saturating signal i_{ph}^{max} to the standard deviation of the noise under dark conditions i_{ph}^{min}:

$$\Delta_{dynamic\_range} = 20 \cdot \log \left(\frac{i_{ph}^{max}}{i_{ph}^{min}} \right) \tag{2.4}$$

where i_{ph} represents the photogenerated current at a certain pixel [114].

Supposing that the photogenerated current i_{ph} in direct proportion to the corresponding intensity level in the same pixel, Equation 2.4 can be rewritten as

$$\begin{aligned} \Delta_{dynamic\_range} &= 20 \cdot \log \left(\frac{\rho_s^{max} \cdot E^{max}}{\rho_s^{min} \cdot E^{min}} \right) \\ &\simeq 20 \cdot \log \left(\frac{E^{max}}{E^{min}} \right) \end{aligned} \tag{2.5}$$

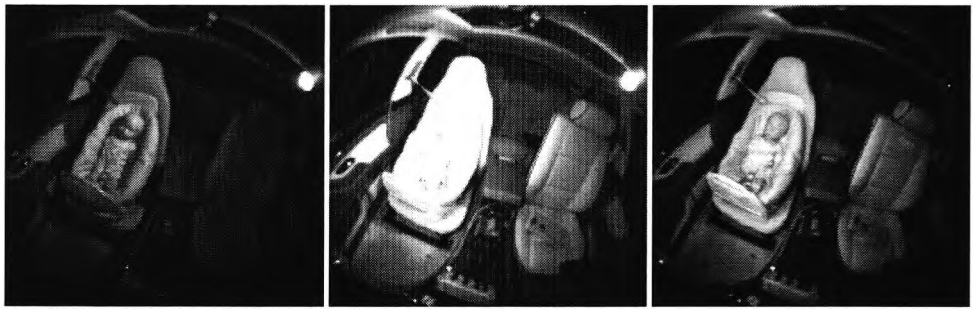
where two surface reflectivity parameters ρ_s^{max} and ρ_s^{min} are assumed to be identical in order to take only the effect of using supplementary illuminations into account. In case that the scene is illuminated with irradiance power of E_{spl}, the dynamic range of the imager becomes

$$\Delta_{dynamic\_range}^{spl} \simeq 20 \cdot \log \left(\frac{E_{amb}^{max} + E_{spl}}{E_{amb}^{min} + E_{spl}} \right) \tag{2.6}$$

Finally, Equation 2.7 shows that the optical dynamic range of a scene with a supplementary illumination $\Delta_{dynamic\_range}^{spl}$ is smaller than a scene only with ambient illumination $\Delta_{dynamic\_range}^{amb}$.

$$\begin{aligned} \Delta_{dynamic\_range}^{spl} &< \Delta_{dynamic\_range}^{amb} \\ 20 \cdot \log \left(\frac{E_{amb}^{max} + E_{spl}}{E_{amb}^{min} + E_{spl}} \right) &< 20 \cdot \log \left(\frac{E_{amb}^{max}}{E_{amb}^{min}} \right) \\ \frac{E_{amb}^{max} + E_{spl}}{E_{amb}^{min} + E_{spl}} &< \frac{E_{amb}^{max}}{E_{amb}^{min}} \\ E_{amb}^{min} \left(E_{amb}^{max} + E_{spl} \right) &< E_{amb}^{max} \left(E_{amb}^{min} + E_{spl} \right) \\ E_{amb}^{min} E_{spl} &< E_{amb}^{max} E_{spl} \\ E_{amb}^{min} &< E_{amb}^{max} \end{aligned} \tag{2.7}$$

qed.



(a) (b) (c)

Figure 2.9: The result of the DoubleFlash method applied to the scene of vehicle interior monitoring: (a) I_{low}, (b) I_{high} and (c) $I_{DoubleFlash}$ with the reduced dynamic range without ambient illumination.

2.3.4 Experimental results

An example of DoubleFlash is presented in Figure 2.9. Two input images I_{low} and I_{high} obtained by a CMOS camera with an optical NIR bandpass filter were illuminated with the different irradiance powers. The details of both the infant and driver seat were not visible in the same image due to the limited dynamic range of the CMOS camera. For example, the vehicle interior in Figure 2.9(a) is not clearly visible while the exposure time of the imager is focused on illuminating the infant seat. On the other hand, the extension of the exposure time of the imager resulted in the over-exposed areas around the infant seat so that all the details of the infant seat was lost. After employing the DoubleFlash method, the image quality for both the infant seat and the vehicle interior was considerably improved as shown in Figure 2.9.

2.4 Shadow removal: ShadowFlash

2.4.1 Motivation

In the field of machine vision, shadows occur frequently in a wide variety of scenes. In many cases, this is undesirable due to the fact that they often lead to the result of irretrievable processing failures. For instance, the shadow cast by an object results in an improper segmentation result with serious artifacts, or detection of an imaginary object. This might result in shadows misclassified as objects or parts of objects due to the over/underestimation in a subsequent matching phase. Accordingly, many existing machine vision algorithms assume that the results of the processing are not under the influence of shadows or that the shadows in an image have been removed [31].

To prevent shadows from being misclassified, they must be explicitly

detected or efficiently removed. Several factors are required to deduce the presence of shadows in a scene: the knowledge of geometric information, the existence of obstructions, and the characteristics of both materials and light sources. Since the knowledge of these factors cannot be readily obtained under real world conditions, it is still a difficult task to identify or eliminate shadows from the scene. Moreover, detecting shadows also involves solving many problems such as region extraction and knowledge representation/integration.

Despite all these difficulties, a number of approaches have been studied to overcome the problem of detecting shadow regions [89, 84, 96, 64]. Existing shadow detection algorithms can be classified in terms of whether the algorithm actively uses knowledge of the environmental conditions or not. The geometric information of a scene and the known directions of light sources are required in identifying shadows in [54]. It also has been shown that shadows can be detected without knowledge of the geometry in an image given the following assumptions [96]: a stationary camera [89], a light source that is strong enough to generate visible shadows, a background containing a sufficient amount of texture, and the dominance of a smooth-shaped background [96]. However, most of these shadow-related algorithms only provide the location of the shadows, and may not provide a complete solution for applications that must suppress the shadows invisibly.

The ShadowFlash method was proposed to solve the problem of removing shadows in an actively illuminated environment by simulating a light source with infinite dimensions [116]. This shadow removal technique was designed to employ the concepts of the DoubleFlash technique introduced in Section 2.3 [52]. Therefore, the ShadowFlash method removes the ambient illumination as well as the cast shadows while the dynamic range of the target scene is compressed. The algorithm requires no boundary extraction task, and it consumes minimal processing time. The idea of ShadowFlash method can be extended into the temporal domain by employing the sliding N-tuple strategy in order to obtain real-time shadowless video sequences without affecting the original frame rate.

This section consists of four subsections. In Subsection 2.4.2, the attributes of shadows and a virtual infinite illuminant plane is analysed. A shadow removal algorithm for video sequences, real-time ShadowFlash is proposed in Subsection 2.4.3 while Subsection 2.4.4 demonstrates the experimental results under various illumination conditions. Finally, the limitations of the ShadowFlash approach as well as the future work are discussed in Subsection 2.4.5.

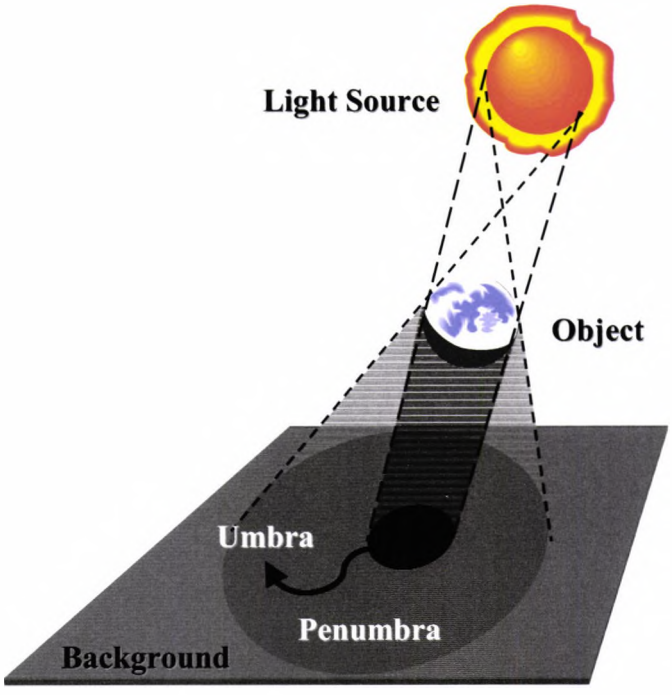


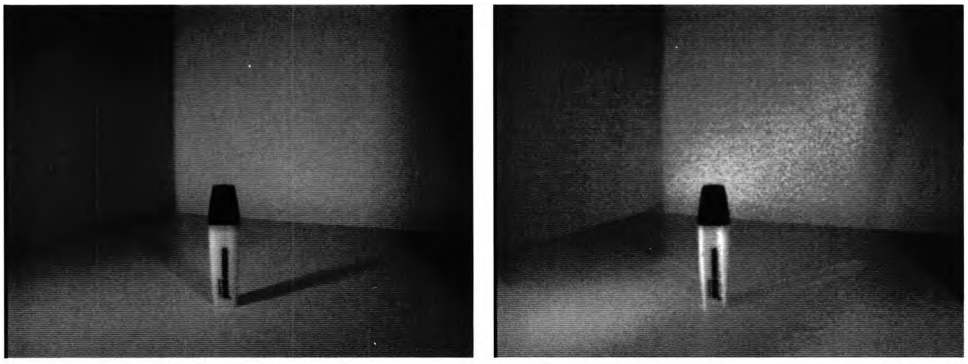
Figure 2.10: Illustration of penumbra and umbra within a shadow scene caused by a spot light source

2.4.2 Analysis

Problem of an attenuated-model based shadow removal approach

Figure 2.10 illustrates the formation of a shadow cast by a bright spot light source. A cast shadow consists of two discernible parts: the umbra and the penumbra. The *penumbra* is a fringe region of half shadow resulting from the partial obstruction of light rays by an object (due to the finite size of the light source), while the *umbra* represents the shadow of the complete obstruction. A narrow penumbra may not appear in an image due to the digitising effects. However, it is not simple to perfectly remove both umbrae and penumbrae when the penumbrae are not negligible in the general image.

The brightness-difference can be estimated through the comparison of the intensity at a pixel in the shadow and the adjacent background, and a weighting factor can be calculated to compensate for the attenuation of illumination within the shadow. The original image and the result of an attenuation-model based shadow removal approach are shown in Figure 2.11. Due to the heterogeneity of the umbra and penumbra regions, the outlines of the shadows are still visible. There are several reliable shadow detection algorithms that identify umbrae and penumbrae separately [28]. However, the attenuation rate of a penumbra is not practically measurable without



(a) (b)

Figure 2.11: Penumbrae problem: (a) a sample image including two shadows caused by two spot lights, and (b) the result of the attenuation-model based shadow removal from (a)

the geometrical knowledge of the illuminating sources.

Infinite illuminant plane

A cloud consists of countless aqueous particles. A light ray passing through the cloud is evenly scattered due to the reflections against the particles. This physical phenomenon causes the photons to be spread over the entire cloud and generate a spatially extended virtual light source. Consequently, on an overcast day, the white sky makes an infinite size of light source, and no shadows occur on the ground (see Figure 2.12).

Based on Gauss's law, it can be proven that the strength of the electric field is independent of the distance from an infinite charged plane. Similarly, the *irradiance*, the amount of light power per surface area, is not influenced by the distance from the light source with infinite extent. It is impossible to build an infinite plane in real life. However, the simulation of an artificial *infinite illumination plane* is possible in the modern computing environment.

2.4.3 Shadow removal

Hypothesis

In case of employing two bright spot light sources, the shadows are classified into four regions as shown in Figure 2.13: the cast shadows influenced by only one light source (a, b), the shadowless region perfectly irradiated by both of the light sources $(a \cap b)$, and the overlapped shadow region only affected by ambient illumination $(a \cup b)^c$.

The formation is interpreted as a Venn diagram as shown in Figure 2.14. Assuming that the universal set represents an image, each set stands for

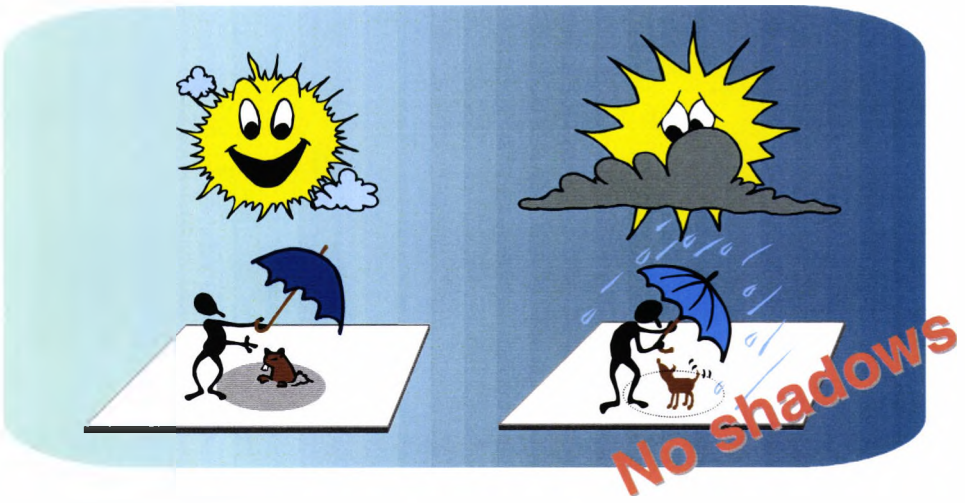


Figure 2.12: Shadows on an overcast day.

an area with constant illumination energy from one light source. The intersection represents the set where the surface is illuminated by two light sources at the same time. Let us denote that the *supplementary* irradiance E_{spl1} which is greater than another supplementary irradiance E_{spl2}, that are emitted by the right and left light sources in Figure 2.13(1) and (2), respectively[6]. Assuming that E_x is the irradiance of the region x, the irradiance map $E(\mathbf{p})$ of an image can be expressed as:

$$E(\mathbf{p}) = \begin{cases} E_a = E_{spl1} + E_{amb} & \text{if } \mathbf{p} \in a \\ E_b = E_{spl2} + E_{amb} & \text{if } \mathbf{p} \in b \\ E_{a \cap b} = (E_{spl1} + E_{spl2}) + E_{amb} & \text{if } \mathbf{p} \in a \cap b \\ E_{(a \cup b)^c} = E_{amb} & \text{if } \mathbf{p} \in (a \cup b)^c \end{cases} \tag{2.8}$$

where $\mathbf{p}$ is a position vector, and E_{amb} represents ambient irradiance.

If the irradiance of all areas ($E_a = E_b = E_{a \cap b} = E_{(a \cup b)^c}$) can be equalised, the simulation of an infinite illuminant plane would be able to deliver the constant illumination power to the entire area as discussed in Subsection 2.4.2. Accordingly, the aim of this work is to equalise the irradiance levels of the area illuminated by active illumination. However, since no information is obtainable to restore the original textures due to the offset reduction scheme, the equalisation task for the area $(a \cup b)^c$ is not considered in our approach ($E_a = E_b = E_{a \cap b} \neq E_{(a \cup b)^c}$).

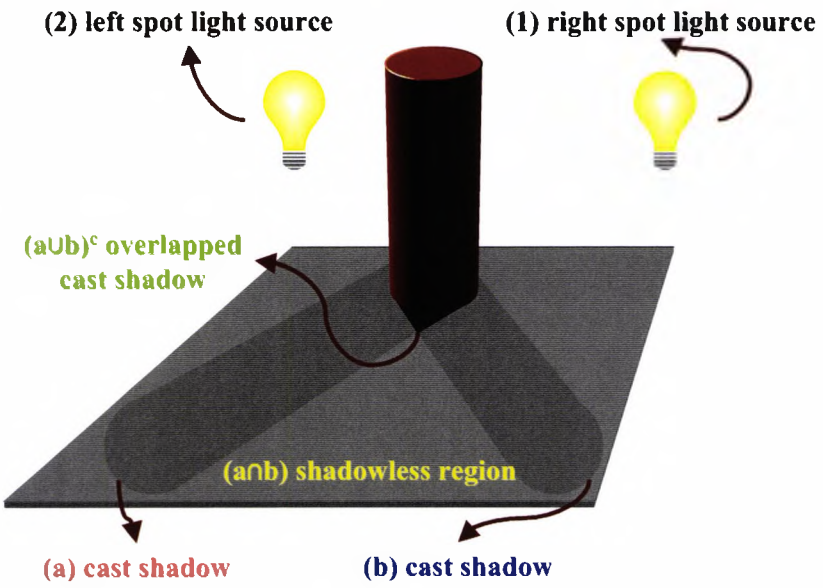


Figure 2.13: Illustration for the formation of shadows with two spot light sources

ShadowFlash

Let us denote the radiant intensity I_x^e as the power emitted from the point light source x into the unit solid angle[7], and the irradiance E_x is the power received at the unit surface element. Assume that the difference of distances between an object and each light sources is a small constant ϵ_x. By neglecting ϵ_x, the relation between the irradiance and the radiant intensity is simplified to

$$E_x = \lim_{\epsilon_x \to 0} \frac{I_x^e}{(d+\epsilon_x)^2} \simeq c \cdot I_x^e$$

where d is the average distance, and c is an adequate constant. Thus, the irradiance power caused by two light sources are equivalent if the radiant intensity of the light source (1) I_1^e is identical with I_2^e (see Equation 2.8 and Figure 2.13).

Assume that three differently illuminated images are used as an input of the system with two separate supplementary light sources. It is supposed that the acquisition time for each image is short enough to neglect scene differences between the input images. For the first image I_a, the left light source has the irradiance E_{spl1} while the right light source has E_{spl2}. And the second image I_b is illuminated with the opposite irradiance to I_a. E_{spl2}

[6]The subscript '*spl*' implies '*supplementary*'.

[7]The radiant intensity in this paper appears with the superscript e *(electromagnetic)* to distinguish it from an image I.

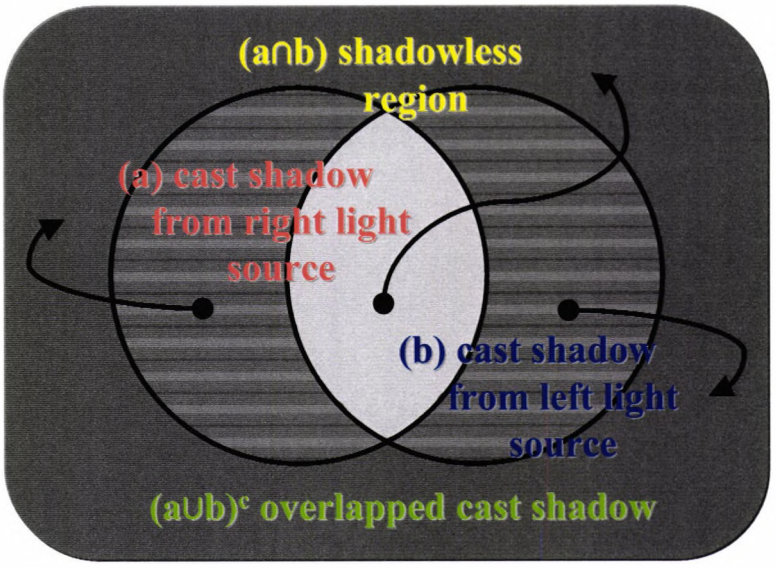


Figure 2.14: A Venn diagram based on the amount of the irradiance power.

is supported to the both sides of the third image I_{offset}. The positions of both light sources are arbitrary, but they must not be coincident.

The distribution of the irradiance for each input image is illustrated in Figure 2.15(a), (b), and (c). Assuming that there is no illumination interference caused by the self-reflection, the supplementary irradiance powers E_{spl1} and E_{spl2} are added to the ambient irradiance E_{amb} while influencing the corresponding parts of the Venn diagram. With the combination of the input images, I_a, I_b, and I_{offset}, one can finally composite an irradiance-equalised image I_{out}. This is given by:

$$I_{out} = |I_a - I_b| + (I_a + I_b) - 2 \cdot I_{offset} \tag{2.9}$$

Assume that the two supplementary illumination sources can illuminate the scene with two different irradiance levels, E_{spl1} and E_{spl2}, and that E_{spl1} is always greater than E_{spl2}. Supposing that $min\,(E_{spl1}, E_{spl2}) > E_{amb}$, then the irradiance of each region is adjusted as:

$$\begin{aligned} E_{a,out} &= |(E_{spl1} + E_{amb}) - (E_{spl2} + E_{amb})| \\ &\quad + \{(E_{spl1} + E_{amb}) + (E_{spl2} + E_{amb})\} \\ &\quad - 2 \cdot (E_{spl2} + E_{amb}) \\ &= \mathbf{2 \cdot (E_{spl1} - E_{spl2})} \end{aligned} \tag{2.10}$$

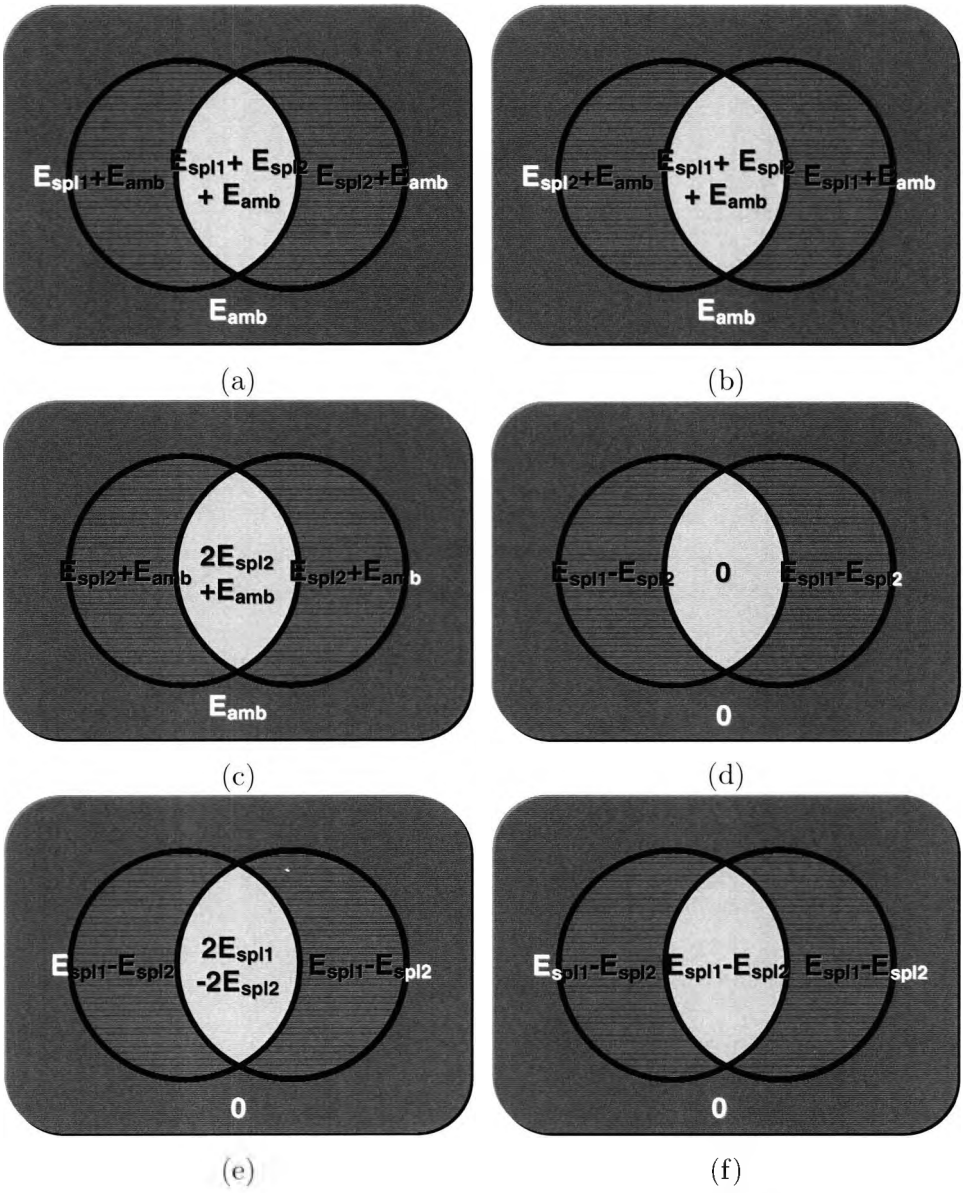


Figure 2.15: Illustration of the shadow removal procedure: (a) I_a, (b) I_b, (c) I_{offset}, (d) $I_{|a-b|} = |I_a - I_b|$, (e) $I_{a+b} = I_a + I_b - 2 \cdot I_{offset}$, and (f) $\frac{I_{out}}{2} = I_{|a-b|} + I_{a+b} - 2 \cdot I_{offset}$

$$\begin{aligned} E_{b,out} &= |(E_{spl2} + E_{amb}) - (E_{spl1} + E_{amb})| \\ &\quad + \{(E_{spl2} + E_{amb}) + (E_{spl1} + E_{amb})\} \\ &\quad - 2 \cdot (E_{spl2} + E_{amb}) \\ &= \mathbf{2} \cdot (\mathbf{E_{spl1}} - \mathbf{E_{spl2}}) \end{aligned} \tag{2.11}$$

$$\begin{aligned} E_{a \cap b,out} &= |(E_{spl1} + E_{spl2} + E_{amb}) \\ &\quad - (E_{spl2} + E_{spl1} + E_{amb})| \\ &\quad + \{(E_{spl1} + E_{spl2} + E_{amb}) \\ &\quad + (E_{spl2} + E_{spl1} + E_{amb})\} \\ &\quad - 2 \cdot (2 \cdot E_{spl2} + E_{amb}) \\ &= \mathbf{2} \cdot (\mathbf{E_{spl1}} - \mathbf{E_{spl2}}) \end{aligned} \tag{2.12}$$

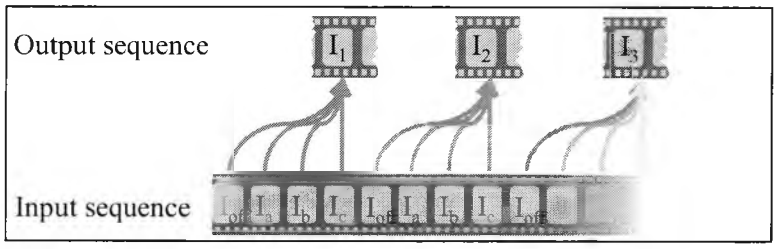
Consequently, the entire region of interest has the same irradiance power as if the image is illuminated by an infinite illumination plane. These regions reconstruct a modified irradiance map $E'(\mathbf{p})$ with the same irradiance $2 \cdot (E_{spl1} - E_{spl2})$. Note that all of these areas are still brightened by the ambient light so that the average intensity of an image might be considerably disturbed by an illumination change within the environment. By employing the *offset reduction technique*, the ambient irradiance E_{amb} can be completely removed. Given the relationship between the irradiance and image as shown in Equation 2.8, an output image without shadows is achieved from the modified irradiance map $E'(\mathbf{p})$. The dynamic range of this output becomes wider in proportion to the difference between two supplementary illumination levels $(E_{spl1} - E_{spl2})$. Figure 2.15(d) and (e) show the procedure of the algorithm, and the resultant irradiance map is illustrated in Figure 2.15(f). In case where one irradiance level of the supplementary light source is zero ($E_{spl2} = 0$), the algorithm can be mathematically replaced by the *max-operation* of the first and second input images $(max\,(I_a, I_b))$.

The number of necessary input frames N_{in} to create one shadow-free image is equal to the number of employed light sources n_{light} plus an additional image for calculating the ambient light suppression.

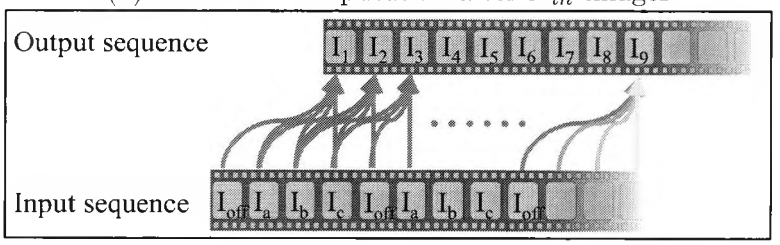
$$N_{in} = n_{light} + 1 \tag{2.13}$$

All the experiments in this work are performed with *three light sources*[8], making the number of inputs four, including ambient light. If the ambient illumination image I_{offset} is negligible, the number of input images can be reduced to n_{light} by ignoring the DoubleFlash. However, the robustness to deal with illumination change is lost.

[8]The minimum number for the practical photometric stereo method. Discussed in Chapter 4.



(a) successive computation after N_{in} images



(b) continuous calculation by the sliding N-tuple strategy

Figure 2.16: Comparison between the non-sliding and sliding N-tuple strategy with three illumination sources

Real-time ShadowFlash: sliding N-tuple strategy

The idea of ShadowFlash can be extended to the temporal domain by synchronising the illumination sources with the trigger signal of a imager so that the imager produces a video sequence of $(\cdots, I_b, I_{offset}, I_a, I_b, I_{offset}, I_a, \cdots)$ where I_x are the images illuminated by the light source x while I_{offset} represents an image having only ambient illumination. However, the direct application of the ShadowFlash method to the temporal domain raises two problems. First, the frame rate of the output sequence will be reduced to $\frac{1}{N_{in}}$ accompanied with a n_{light}-frame delay in the beginning of the acquisition, because N_{in} images are required to obtain one shadowless image as explained in Equation 2.13. Secondly, if any object in the scene moves during a N_{in}-tuple, some artifacts will occur around the boundary of the object.

In order to avoid the frame rate reduction, a *sliding N-tuple strategy* is proposed. A memory window with the width of N_{in} frames is created, whereby the window is moving along the time axis. In the window, N_{in} differently illuminated successive images are constantly refreshed. These images continuously form a set of inputs to create a shadow-free output image. In Figure 2.16 an example of the sliding N-tuple method for three supplementary light sources ($N_{in} = 4$) is given. Figure 2.16(a) shows that the frame rate of the result sequence is divided by *four* while the output frames are consecutively calculated by employing the sliding N-tuple strategy in Figure 2.16(b).

Established shadow detecting algorithms commonly employ a set of spa-

tial convolution filters for detecting image regions with damped texture compared to a shadow-free reference image. Hence, the number of necessary fixed/floating point operations per pixel are proportional to the number of elements within the spatial convolution masks. One of the most important advantage of the proposed algorithm compared to conventional shadow detection approaches is the processing cost. The proposed algorithm requires only a few fixed point operations per pixel for both detecting and removing shadows. Another advantage is that no shadow-free reference image is necessary which has to be updated over time. Finally, since the ShadowFlash is the extended version of the DoubleFlash technique, it also compresses the dynamic range of the scene while the ambient illumination is completely removed.

Fast moving objects may distort the result of the sliding N-tuple strategy. The amount of distortion depends on the frame rate of the imager. When the imager produces frames with sufficient speed, the artifacts caused by moving objects should be negligible. In case of a slow frame rate compared to the velocity of moving objects within the scene, a supplementary algorithm should be implemented to detect and correct the difference between frames. However, if such a correction filter is added to the ShadowFlash approach, the speed advantage over the other algorithms will be reduced or lost.

2.4.4 Experimental results

Several different experiments to demonstrate both the basic idea of the ShadowFlash and sliding N-tuple strategy are conducted. In the first experiment, two identical halogen bulbs are used for supplementary illuminations, while another halogen lamp is installed for simulating ambient illumination. The irradiance power E_{spl2} is minimised in order to maximise the dynamic range of the output image. A CCD camera is used for taking images with 640×480 pixel resolution in 8-bit intensity levels. The positions of both the bulbs and camera are chosen to minimise the overlapped shadow regions. The experiment is performed with a metallic object on the complex-textured background.

Figure 2.17(a) and (b) represent the input images with the existence of both the supplementary and ambient illuminations from different directions. Some parts of the texture on the background are obscured due to the shadows, although the textures are still visible within them. The histograms of these images are also shown in Figure 2.18(a) and (b), respectively. In Figure 2.17(c), the image illuminated only by the ambient light is shown.

The results of the interim stage of the procedure are shown in Figure 2.17(d) and (e). Figure 2.17(d) shows the composite image of the two input images with supplementary illumination. In this step, the intensity

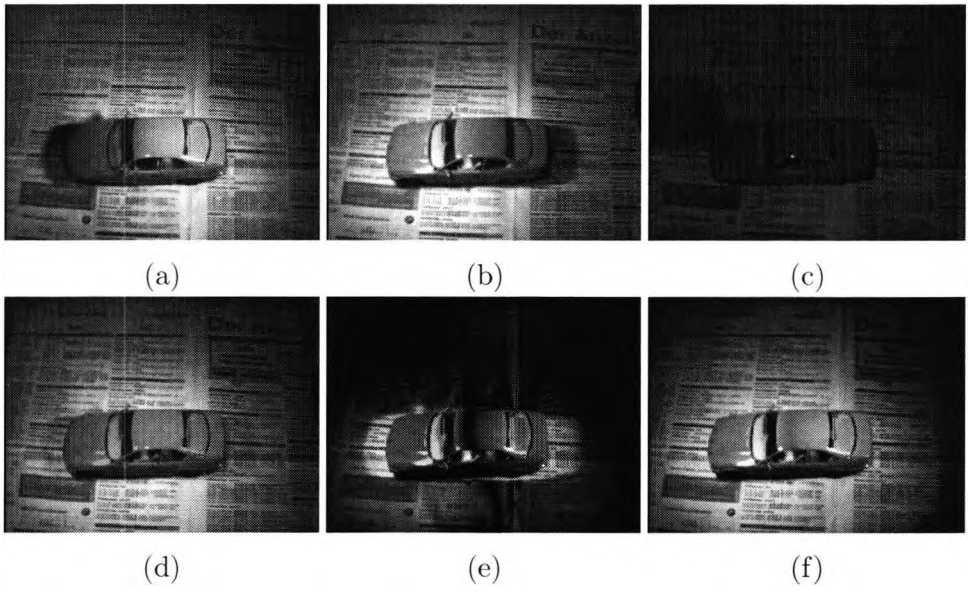


Figure 2.17: Examples of ShadowFlash with the ambient illumination: (a) input image with the right light source I_a, (b) input image with the left light source I_b, (c) input image only with the ambient illumination I_{offset}, (d) $I_a + I_b$, (e) $|I_a - I_b|$, and (f) the result of ShadowFlash algorithm I_{out}

resolution of the image is temporarily doubled to 9 bits due to the addition process $(I_a + I_b)$ as shown in Figure 2.18(d). In principle, all of the textures in the input images are identical. Thus, it is obvious that the pixels which have the intensities greater than zero after the subtraction process $|I_a - I_b|$ have been illuminated with different irradiance powers (see Figure 2.17(e)). Figure 2.18(e) shows the distribution of the histogram biased to zero.

Figure 2.17(f) shows the result image of the ShadowFlash algorithm. The shadows which have covered the background textures are successfully removed while the patterns of the background are completely restored by simulating the illumination from an infinite illuminant plane. The dynamic range of the field of view is also conserved as shown in Figure 2.18(f). The intensity resolution of the image has doubled in the result of the addition phase. However, the frequency per every 2nd intensity level has zero value because another addition-operation of the algorithm makes all the intensity values turn into even numbers in the final task as shown in the small window in Figure 2.18(f). Therefore, the intensity resolution can be compressed to 8 bits again by eliminating the lowest bit.

Another experimental result for colour images is shown in Figure 2.19. The input images are taken by a colour CCD camera having both the Auto Gain Control (AGC) and Gamma correction functions enabled. Since the shadows cast by the ambient illumination in the input images are not visible or very weak due to the effect of the nonlinear intensity compression, the

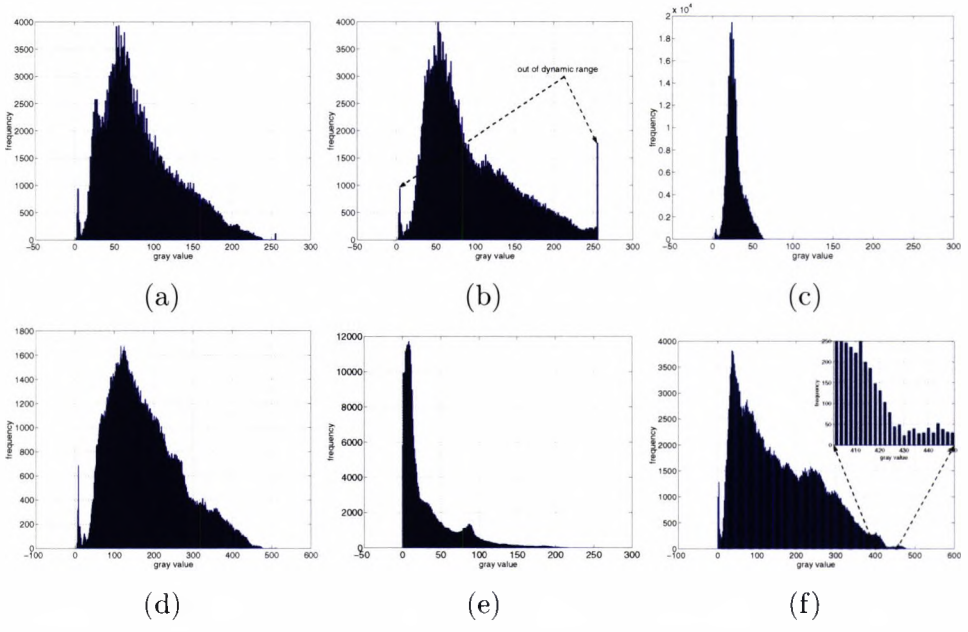


Figure 2.18: Intensity histograms for the inputs and the result where the ambient light exists: (a) I_a, (b) I_b, (c) I_{offset}, (d) $I_a + I_b$, (e) $|I_a - I_b|$, and (f) I_{out}

algorithm could be modified to:

$$I_{out} = |I_a - I_b| + (I_a + I_b) = max(I_a, I_b) \tag{2.14}$$

Consequently, the influence of ambient illumination is not suppressed in I_{out}.

An example for outdoor images is presented in Figure 3.7. There is a considerable amount of time interval between the scenes in Figure 3.7(a) and (b). Since the ambient illumination information I_{offset} is not available, the simplified algorithm shown in Equation 2.4.4 is used. Although most of the shadows are successfully eliminated, some artifacts emerged in the result because of the scene difference such as the moving pedestrians.

In Figure 2.21, the shadows are not completely removed because the irradiance powers of the supplementary illuminations are not evenly distributed over the field of view. The self-reflection caused by the surface with a large constant reflectance is another reason. The overlapped shadow between two objects (the shadow of the right side-view mirror) still remains due to the limitation of the proposed algorithm.

Finally, a sequence synchronised with three infrared illumination sources installed in different positions and the results of real-time ShadowFlash are shown in Figure 2.22. The scene contains a background with complex textures while ambient illumination exists. The sequence is recorded by a HDR

(a) (b) (c)

Figure 2.19: Examples of ShadowFlash for colour images taken by a CCD camera with AGC: (a) I_a, (b) I_b, and (c) I_{out}

(a) (b) (c)

Figure 2.20: Examples of ShadowFlash for outdoor images:(a) I_a, (b) I_b, and (c) I_{out}

CMOS camera at 30 frames per second with 324×244 pixel resolution and 12-bit intensity levels.

In the original sequence, the ambient illumination is coming through the windows and therefore the background is dimly visible. Each frame illuminated by the installed light sources contains at least one strong cast shadow. The results are very successful since all the cast shadows are completely removed while the ambient illumination coming through the windows is suppressed. The complex texture of the background is also satisfactorily restored. Some visible distortions occur around the moving object when the object moves fast. However, the influence of those artifacts may not be significant if the results are applied for object tracking purpose.

2.4.5 Discussion

A real-time shadow removal method based on the radiation power analysis is proposed. With a reasonable number of controllable supplementary illuminations, the ShadowFlash algorithm simulates an infinite illumination plane over the field of view and eliminates both shadows and ambient illuminations from the scene. By employing the sliding N-tuple strategy, the idea could be extended to the temporal domain. Finally, the proposed approach could successfully remove cast shadows from a complex-textured

(a) (b) (c)

Figure 2.21: Less successful case due to the uneven distributed illumination: (a) I_a, (b) I_b, and (c) I_{out}

scene without distorting the recovered background. Another achievement is that the algorithm works without any support of region extraction tasks, so that its processing time is dramatically decreased compared to the other shadow detection algorithms. High reliability can be also guaranteed if the irradiance powers of active illuminations are evenly distributed.

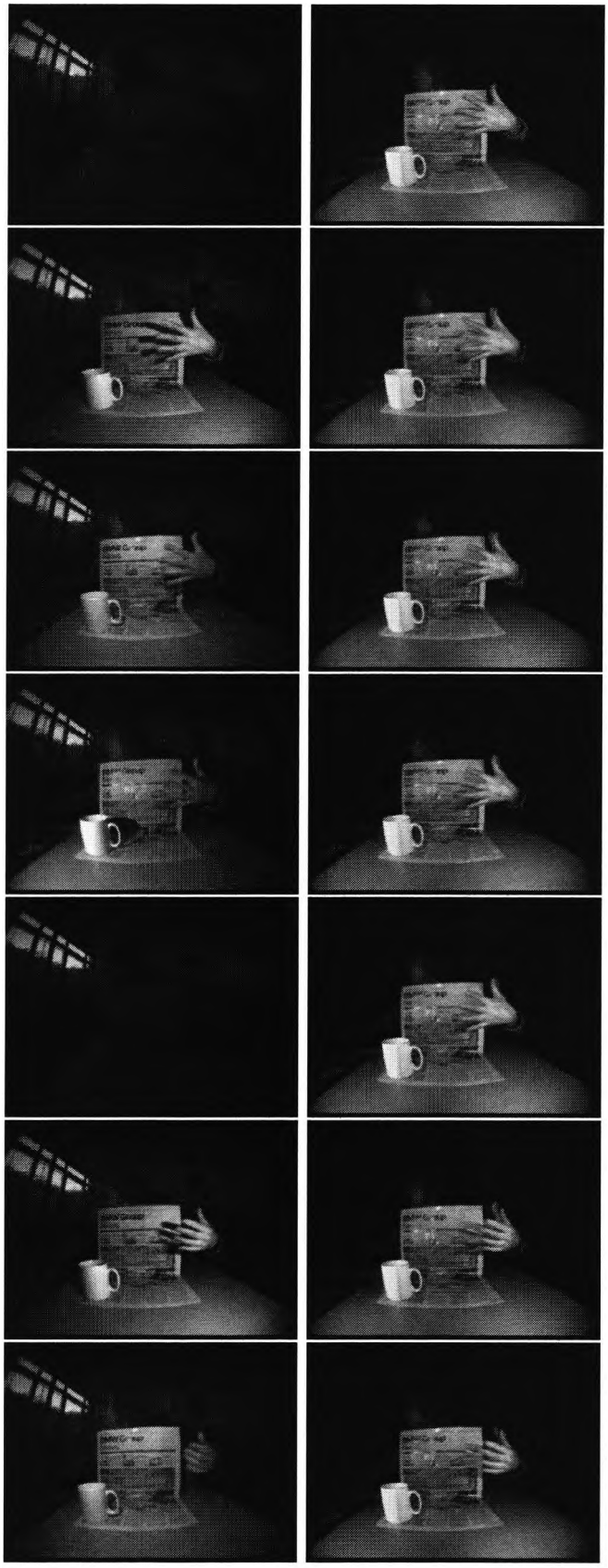
Limitations and required conditions

Several requirements had to be met in order to obtain the satisfactory results in the experiments.

- The dynamic range of the imager must be wide enough to cover the entire local dynamic range of the scene. i.e. there must not be any pixels which are over-exposed in an input image.
- The irradiance on the target surface from each light source should be equal to obtain satisfactory results. i.e. the radiant intensities I^e of the light sources should be approximately equivalent while the distance d from each light source to the target surface is similar.
- The self-reflection on the surface of an object caused by the supplementary light sources could result in some artifacts upon the recovered background. To avoid the undesirable effects, the supplementary illuminations should be carefully positioned taking both the camera and object into account.
- The frame rate must be sufficiently fast for minimising the ambiguity caused by motion.
- Finally, the overlapped shadow region must be minimised.

2.5 Precis

In this chapter, various problems caused by the limited dynamic range of the present imaging sensors were discussed. After defining the high dynamic range environments, a few hardware-level solutions based on manipu-



(a) (b)

Figure 2.22: Sample sequence of Real-time ShadowFlash: (a) original sequence and (b) Real-time ShadowFlash

lating sensor responses to extend the dynamic range of the imaging sensors were introduced.

Finally, two image processing based approaches for improving both the optical dynamic range and the quality of an image acquired by a conventional imager with support of active illumination were discussed.

DoubleFlash Mainstream CCD based and most of the emerging CMOS based image sensors do not provide sufficient optical dynamic range for monitoring the interior of a vehicle where people experience extreme variations of illumination either spatially or temporally [51]. In order to capture images without losing image details in such an environment, it is essential to employ an imager suitable to the high dynamic range and/or a novel approach to decrease the dynamic range without varying illumination offset. The *DoubleFlash* technique was employed in the proposed system, which combines the advantages of offset reduction and dynamic range compression by illuminating two input images with different radiant intensities, originally introduced in [52].

ShadowFlash Nearly all vehicle interior monitoring applications introduce supplementary light sources (usually in the near-infrared region) in order to attain an appropriate illumination offset. Therefore, strong cast shadows are unavoidable in the field of view. Shadows often create erroneous segmentations causing false detection of imaginary objects, which hinders the overall performance of a system. The *ShadowFlash* method is a method to eliminate shadows without distorting the original textures of the scene by simulating a virtual light source of infinite size. The algorithm uses multiple images where each image has been flashed from a different direction. The number of necessary input images N_{in} to create one shadow-free image is equal to the number of employed light sources n_{light} plus an additional image for calculating the ambient light suppression. Since the proposed approach does not require any region extracting tasks, an advantage in perspective of lower computing cost and processing reliability could be achieved compared to the conventional shadow detection/removal algorithms.

Chapter 3

2D processing: object segmentation

3.1 Motivation 46
3.1.1 Introduction 46
3.1.2 Segmentation for a vehicle cabin environment . . . 48
3.1.3 Overview 49
3.2 Approximate boundary extraction 50
3.2.1 Texture-based object detection 50
3.2.2 Morphological operations 52
3.2.3 Deciphering of object of interest 53
3.3 Active contour models 53
3.3.1 Introduction 53
3.3.2 Fundamentals 54
3.3.3 Dynamic programming 57
3.3.4 Convexity defects driven active contour models . . 60
3.4 Experimental results 61
3.5 Precis 63

3.1 Motivation

3.1.1 Introduction

Segmentation is an activity to obtain a compact representation for distinguishing objects of interest from a background. For most vision applications, segmentation is a key step in image analysis. For example, the detection of the numberplate position is critical for an automatic numberplate identification system. Separating individual characters from the words is a crucial task for analysing a document in the field of handwriting recognition. For occupant detection systems, the occupant should be distinguished from the irrelevant background components such as a passenger seat.

Autonomous segmentation is one of the most difficult tasks in image processing, and this step often determines the eventual success or failure of the overall system operation. False extraction of an object boundary may result in illusory objects causing misjudgement on analysing the object and decrease the reliability as well as the system performance. Effective segmentation techniques are critical to a successful solution, therefore, considerable care should be taken to design a segmentation process.

Most of early segmentation algorithms are based on one of two basic properties of intensity values: *discontinuity* and *similarity* [30]. Segmentation approaches based on these properties are well studied in the last decades and appear commonly in various vision applications [43, 30. 95, 24]. In the first category, the approach is to partition an image by analysing the attributes of an image such as abrupt changes of intensity level. The principal areas of interest within this category are observation of lines or edges of the object of interest. However, a method based on the edge property is prone to failure due to the weak capability of connecting broken lines in the presence of blurring. *Similarity* is another key property for segmentation. A good similarity measurement usually provides more effective segmentation results than the approaches based only on the discontinuity property. The principal approaches in this category are based on thresholding, region growing, as well as region splitting and merging. For example, a thresholding technique which makes a decision based on local pixel information could be effective if the foreground intensity level is sufficiently different from the range of the background intensity level.

Although the segmentation algorithms based on elementary image attributes are relatively straightforward to implement, their limited robustness to noise makes it difficult to derive reliable segmentation results. Recently, most studies in the field of segmentation focus on developing algorithms capable to overcome problems which could not be handled by the conventional approaches. The followings are the descriptions of three most popular approaches in modern segmentation research: *region-based approaches*,

probabilistic models and *elastic models*.

Region based approaches A region-based method is based on the assumption that the goal of segmentation is to determine which components of a data set naturally belong together. An image is partitioned into connected regions by grouping neighbouring pixels of similar intensity levels. Adjacent regions are then merged under criteria such as homogeneity or sharpness of region boundaries. Overstringent criteria create unwanted image fragments while lenient ones overlook blurred boundaries and overmerge. Hybrid techniques using a mixture of these two methods are also popular. A good example is *clustering*, originally introduced by Ohlander [70]. More sophisticated models based on various statistical characteristics of an image have been utilised as clustering methods [87, 105, 111]. The partition based on *normalised cuts* introduced by Shi [92, 93] have also shown some success. The normalised cuts split an image into homogeneous groups by minimising the disassociation between the groups and maximising the association within the group. Although the region-based methods are quite useful for particular applications, most of these approaches tend to be rather *arbitrary* since there is not much theory available to predict what/how should be clustered at the end of segmentation process.

Probabilistic models A number of important vision problems could be phrased as problems of missing useful elements of the data. These problems are addressed by this missing variable model. Segmentation could be thought of as a method to determine which of a number of sources a measurement came from. For example, segmenting an image into regions involves determining which source of color and texture generated the image pixels (i.e. which region a pixel belongs to in the sample).

Most of the missing variable models applied to the image segmentation are based on mixture models. The basic assumption underlying the segmentation is that the different image layers present in a pixel contribute independently to its intensity. Therefore, the intensity of a pixel is the sum of the brightness of image layers which compose that pixel. A successful inference algorithm, known as *expectation maximisation* (EM) [20], can be used to compute maximum likelihood estimates given incomplete samples for various segmentation models. A number of studies have been made for the segmentation model selection, especially for motion and ranged data [63, 47, 8].

The standard problem in segmentation using missing variable models is to predict the number of missing variables (image layers) beforehand. This is particularly difficult for a scene where the textural attribute of the object of interest is ambiguous and no geometric constraints

are given. Since most of the approaches for solving the missing variable problems involve intensive iteration procedures, segmentation by probabilistic models is less suitable for applications operated in an embedded real-time system with limited hardware capability.

Elastic models A connectivity-preserving relaxation-based segmentation method, usually referred to as either the *active contour model* or *snake*, was first introduced by Kass in [46]. The method starts with some initial boundary shape represented in the form of spline curves, and iteratively modify it by applying various shrink/expansion operations as some energy functions are minimised. The energy functions generally consist of (i) internal contour force which enforces the smoothness, (ii) image force which attracts the contour to the desired features, and (iii) external constraint force.

This active contour model provides a powerful interactive tool for image segmentation. However, since this approach relies on strictly local information, the original snake model is vulnerable to image noise. The preliminary shape of a target object must be given before active contour proceeds with its evolution. Thus, the active contours are especially useful when either *a priori* information of the target object is given, or at least the approximate boundary of the object of interest is predictable.

Despite of numerous efforts for decomposing an image into useful groups, there are no comprehensive theories of segmentation at the moment. Since segmentation remains an open problem in vision, the key issue is to determine what representation is suitable for the problem at hand.

3.1.2 Segmentation for a vehicle cabin environment

For the vehicle interior monitoring applications, few approaches have employed segmentation techniques [55, 62, 50, 101]. Since most of these studies are based on binocular stereo vision techniques, a segmentation task is not essential for recovering object surfaces. Nevertheless, valuable information could be extracted by analysing two-dimensional geometry of the target object. In [51], Koch showed that the vehicle occupant classification result could reach 95% by utilising only *two-dimensional geometric information* obtained from a single monochrome imager.

The difficulty of video object segmentation mainly comes from the inconsistent *deformation* of objects. In case of a *non-rigid object*, the segmentation task becomes even more challenging due to the diversity of shapes originating from its unpredictable deformation. Incidentally, some occupant classes for the vehicle in-cabin monitoring involve highly unpredictable movements as well as frequent exaggerated deformation properties.

Some active stereo vision techniques such as the *photometric stereo method* require integrating surface normal vectors for reconstructing the object surface. In this case, the boundary extraction process is indispensable for successful results. Assuming that the intensity at a part of a target object is overwhelmed by noise, this could propagate unacceptable errors through the vector estimation, and unrecoverable distortions may be produced on the entire surface of the object as the result of vector integration. Since most vector integration methods are not capable of handling abrupt changes of depth, it is also important to isolate an area with closer ranges of depth.

Despite the importance of the aforementioned problems, surprisingly, little of the research related to vehicle interior monitoring has paid much attention to developing a segmentation strategy. For example, Koch proposed a segmentation method based on the similarity analysis between frame sequences in [51]. Although the method provides strong edges when objects have prominent motions, the unreliable performance for stationary objects makes it difficult to employ the method directly to the active vision techniques as discussed above.

In this section, an active contour model based on concavity analysis is proposed to solve the above-mentioned difficulties. The main reasons for employing the active contour model for this work are: (i) the ease of initialising an active contour model due to the fact that the object of interest is supposed to dominate the field of view, (ii) the ability to handle diverse shapes caused by unpredictable deformations and finally (iii) the reliability for providing precise object boundaries in most circumstances. The improved boundary should facilitate the surface reconstruction by discarding unnecessary and/or inaccurate information. On the other hand, since motion estimation is very time-consuming and usually unreliable for non-rigid objects, the analysis of motion is not considered in the proposed approach.

3.1.3 Overview

Like most of other machine vision applications, the boundary extraction task is of great importance in the proposed system to provide useful primary information. In this work, the textural similarity of an input frame with respect to a *reference background* is computed based on their local statistical properties. While the local and global illumination changes are stabilised by the DoubleFlash technique and all the cast shadows are removed by ShadowFlash, the morphological operations provide an *approximate boundary* of the target object by merging image fragments. The approximate boundary is then used for initialising an *active contour model*. In order to improve the mobility of the active contour, *convexity defects* are computed by investigating the *convex hull* generated around the initial boundary.

3.2 Approximate boundary extraction

3.2.1 Texture-based object detection

To discriminate a target object from a background is an important task. A way to perform this task is to subtract the input image from the *reference image* containing only background components. Suppose that the camera position is stationary and illumination is constant over time, the background subtraction leaves only non-zero areas that correspond to the transitional (foreground) image components. Assuming that an *object image* I_{obj} is a binary image which isolates the foreground components, the I_{obj} could be expressed as a function of the target I_{target} and reference image I_{ref}:

$$\begin{aligned} I_{obj}\left(I_{target}, I_{ref}\right) &= 0 \quad \text{if } f_{comp}\left(I_{target}, I_{ref}\right) \leq \varepsilon \\ &= 1 \quad \text{otherwise} \end{aligned} \tag{3.1}$$

where ε is a small positive number and f_{comp} represents a textural similarity measurement which returns a real number proportional to the closeness between two input images.

Background maintenance

Given an appropriate background, many formidable problems can be resolved easily by separating a foreground from a background. Normally, the difficult part of the background subtraction is not the subtraction task itself but the design of a background model.

In certain situations, only limited space is allowed between an imager and object of interest. For example, a passenger monitoring system usually suffers from the limited physical distance between the installed camera and occupant. In this case, a significant portion of the background is constantly occluded by the target object. Since the change of the background is only partially monitored unless the occupant disappears from the scene, the background maintenance becomes a difficult problem. Especially when the object has either relatively low mobility or no motion, it is extremely difficult to predict the object of interest without the support of a reliable segmentation task. For example, a child seat in a vehicle may not be distinguishable from the movable passenger seat where the child seat is mounted. During a specific time interval, a child seat could generate less motion than a passenger seat. In this case, the child seat could be falsely classified as a part of the background and be erroneously updated along with the real background.

An unstable illumination condition is another factor that makes the background maintenance difficult. The frequent change of ambient illumination disturbs the background updating process by increasing the uncertainty for

estimating brightness at a specific pixel position. An illumination condition with a constant vibration also makes the updated background image ambiguous.

Despite the fact that great efforts have been spent on solving the abovementioned problems [67, 59, 22, 80, 98, 97, 16, 119, 120, 88, 79], presently there is no ultimate background maintenance algorithm which guarantees sufficient reliability for *safety-critical* applications. Although the problem of the illumination fluctuation have been successfully addressed by employing the illumination techniques discussed earlier in Chapter 2, the rest of the problems may not be solved by any vision approaches. Therefore, a *fixed background* taken by a stationary imager is assumed in this work. By subtracting the fixed background from the sequences, *fixed pattern noise* caused by the imager could also be suppressed. Nonetheless, the use of the fixed reference background in reality should be minimised due to some movable interior parts in a vehicle such as passenger seats.

Similarity measurement based on statistical analysis

Statistical analysis of textures involves the computation of the distribution of certain properties such as gray level, average value, deviation, dispersion, entropy, etc. For example, if the histogram of an image segment is divided by total number of pixels of the segment, the result represents the probability that the certain gray level appears in the image segment. A typical deviation shows the dispersion with respect the average value.

It has been shown that any statistics higher than second-order contain little information that could be used for the texture analysis [45], and that region-based image processing normally provides more reliable results than pixel-based approaches. Consequently, a similarity measuring function which investigates *first-* and *second-order statistics* associated with the *subimaging* is employed for detecting an object of interest in this work. The size of the subimage is empirically chosen to be 5×5, which is comparable to the size of a *texture primitive* of the applied scene. The textural similarity function f_{comp} consists of a series of two low-order statistics and a intensity difference between the two input images. Assuming that the intensity difference image I_{diff} is derived from subtracting two input images I_{target} and I_{ref}, the similarity measurement function $f_{comp}(\cdot)$ is defined as

$$f_{comp}(I_{target}, I_{ref}) = \alpha \cdot \mu_{diff} + \beta \cdot \sigma_{diff} + \gamma \cdot I_{diff} \qquad (3.2)$$

where μ_{diff} and σ_{diff} are the first- and second order statistics of the difference image and α, β and γ are weighting constants. The first and third terms of the similarity measurement function might be replaced with more sophisticated filters such as a Gaussian filter.

3.2.2 Morphological operations

A thresholding process often separates an object into scattered image pieces when the object has either complex textures or extreme contrast. This occurs especially when the average intensity level of the foreground is similar to the one of the background, and may lead to the failure of subsequent processes by falsely discarding critical information. The problem could be resolved by combining together image fragments in close proximity which originally belong to a single object.

Morphology is a technique of image processing based on shapes [33]. The basic idea in mathematical morphology is to convolve an image with a given mask known as a *structuring element*, and to binarise the result of the convolution using a given function. The structuring element is a binary matrix used to define a neighbourhood shape and size for morphological operations. It consists of only ones and zeroes which define an arbitrary shape and size. By choosing the size and shape of the structuring element, a morphological operation sensitive to specific shapes can be constructed. Functional descriptions of two principal morphological operations are as follows [102]:

Dilation Dilation adds pixels to the boundary of an object in an image. The dilation process is performed by laying the structuring element on the image and sliding it across the image in a manner similar to convolution. If the origin of the structuring element coincides with a '0' pixel in the image, there is no change; move to the next pixel. If the origin of the structuring element coincides with a '1' pixel in the image, perform the OR logic operation on all pixels within the structuring element. With a dilation operation, all the '1' pixels in the original image will be retained while any boundaries will be expanded and small holes will be filled.

Erosion Erosion removes pixels on object boundaries. The erosion process is similar to dilation, but the operation turns pixels to '0', not '1'. As sliding the structuring element across the image, there is no change if the origin of the structuring element coincides with a '0' in the image. If the origin of the structuring element coincides with a '1' in the image, and any of the '1' pixels in the structuring element extend beyond the object ('1' pixels) in the image, then change the '1' pixel in the image to a '0'.

Many complex operations can be also defined based on various combinations of multiple applications using dilation and erosion. The most useful of these for morphological filtering are called *opening* and *closing*. Opening consists of erosions followed by dilations. This cleans up the image by removing small bright spikes or noise and then returning the remaining objects to their original size. Reversely, closing is formed by dilations followed

by erosions and is used to fill small holes in an object and/or to join broken boundaries into continuous segments.

Binary morphology has been successfully used in several segmentation systems [90, 78, 115, 56]. In this work, the closing operation is used to eliminate small noise particles. To recombine separated image blobs, dilation operations are repeatedly performed *ten* times. Finally, an approximate boundary is generated which constantly encloses the object of interest. For these operations, a 3×3 *rectangular* structuring element is used since the shape of the structuring element does not significantly influence system performance.

3.2.3 Deciphering of object of interest

Even after the morphological operations, multiple image fragments can remain unconnected, and a decision should be made for selecting the object of interest among the image blobs. In many cases, reliable determination of an object of interest is only possible after significant information has been extracted about the object portrayed. Unfortunately, methods for extracting this information often require a shape description which could be provided only by a recogniser. This is of course not optimal since the overhead of a recognition task usually exceeds that of segmentation itself. For example, the segmentation of a particular person in a group of people may not be possible without the supervision of a recogniser specialised to that person.

For some applications where an imager is placed close to an object of interest, the solution to this problem becomes much simpler. In case of vehicle interior monitoring applications, the camera location is constrained by the limited physical space inside a vehicle. This usually causes the object of interest to dominate the field of view. In this work, the *largest* image blob is chosen as an target object in case that the size of the blob is greater than a pre-determined threshold. The two-dimensional location of the object centroid with respect to the centre of the camera view is also used since the object of interest is usually located closer to the camera focus than other unimportant objects.

3.3 Active contour models

3.3.1 Introduction

An active contour model called a *snake* is an energy-minimising spline guided by external image and internal spline forces. Snakes may be understood as a special case of a more general technique of approximating a deformable model to an image by means of energy minimisation, while the mobility of

each snaxel[1] depends on the contour's shape and location with respect to the target object [95].

Numerous provisions have been made in the literature to improve the robustness and stability of the snakes [1, 29, 14, 58, 10, 69, 9, 76, 77, 12, 91, 103]. For example, Cohen introduced a *balloon force* which can either inflate or deflate the contour [14]. This force helps the snake to ignore spurious isolated weak image edges and counters its tendency to shrink. The snake with the balloon force becomes more robust to the initial position and image noise. Nevertheless human intervention is still necessary to decide whether an inflationary or deflationary force is needed. Amini suggested using dynamic programming to minimise the energy function [1]. This method exhaustively searches all the admissible solutions, and each iteration results in a locally optimum contour. Geiger proposed to solve the problem in a single iteration by allowing the contour to be searched from anywhere in a large area around the initialisation position [29]. Neuenschwander proposed to let the user specify the two end points of the desired contour [69]. As the optimisation process progresses, the edge information is propagated from the end points towards the centre. Fua proposed to reach the desired goal by imposing attractor and tangent constraints, where the attractor constraint forces the contour to move towards or pass by a particular point in the image while the tangent constraint forces the contour to have a certain tangent at a particular point [27]. Finally, the concept of active contour has been successfully extended to perform tasks such as edge and subjective contour detection, motion tracking, stereo matching and image segmentation [107, 83, 113, 121, 60].

In this section, an active contour model based on the *convexity defects* analysis is proposed. The approximate boundary produced by the tasks described in Section 3.2 is used for initialising the active contour. A *convex hull* is created around the approximate boundary and used for computing the convexity defects of the contour. The convexity defect improves the deformation of the active contour by providing higher mobility at a concave boundary.

3.3.2 Fundamentals

The active contour model used in this work was originally introduced by Atkins in [2]. An active contour V is defined as a collection of n-snaxels in the two-dimensional coordinates:

$$V = \{v_1, v_2 \cdots, v_n\} \qquad \text{where } v_i = [x_i, y_i] \tag{3.3}$$

[1] A snake cell: for the rest of this work *snaxel* will be used to refer to such points or elements of the snake.

The snake model is basically an elastic curve which can dynamically conform to object shapes in response to internal force E_{int} and external force E_{ext}. Given the contour V. an energy function for the contour can be stated as

$$E(V) = \sum_{i=0}^{n} (\alpha \cdot E_{int}(v_i) + \beta \cdot E_{ext}(v_i)) \tag{3.4}$$

where the internal forces E_{int} keep the snake smooth and the external forces E_{ext} couple the snake to the target image attracting the snake to features of interest. For segmentation applications, these features could be object boundaries. The α and β are weighting constants.

The internal and external forces (E_{int} and E_{ext}) are matrices where the value at the center of each matrix corresponds to the contour energy at the i-th snaxel v_i. Other values in the matrices correspond spatially to the energy at each point in the neighbourhood of v_i. The snaxel v_i is moved to the point v_i' corresponding to the location of the minimum value in the neighbourhood matrix. This process is illustrated in Figure 3.1. Finally the snake contour V approaches to the boundary of the target object as the energy function E is minimised and stops its deformation when this minimisation process reaches the minima.

Internal Energy

The internal energy function is intended to enforce a shape on the deformable contour and to maintain a constant distance between the snaxels in the contours. Additional sub-functions can be added to influence the motion of the contour [3, 76, 77]. The internal energy function E_{int} is defined as follows:

$$\alpha E_{int}(v_i) = cE_{coninuity}(v_i) + bE_{balloon}(v_i) \tag{3.5}$$

where $E_{continuity}(\mathbf{v}_i)$ is the *continuity energy* which compels the snake to have a smooth shape while $E_{balloon}(\mathbf{v}_i)$ is a *balloon force* causing the polygon to expand or contrast. The constants c and b provide the relative weighting of the energy functions.

Continuity Energy In the absence of other influences, the continuity energy forces a closed deformable contour to form a circle. Each energy element $e_{jk}(v_i)$ in the matrix $E_{continuity}$ is defined as follows:

$$e_{jk}(v_i) = \frac{1}{\|V\|} \|n_{jk}(v_i) - \gamma(v_{i-1} + v_{i+1})\|^2 \tag{3.6}$$

where $n_{jk}(v_i)$ is one of the snaxel candidates at the coordinate (j, k) in the neighbourhood matrix. For a closed contour, the contour V is given a modulus of n. i.e. $v_{n+i} = v_i$. γ is then defined as follows:

$$\gamma = \frac{1}{2\cos\left(\frac{2\pi}{n}\right)} \tag{3.7}$$

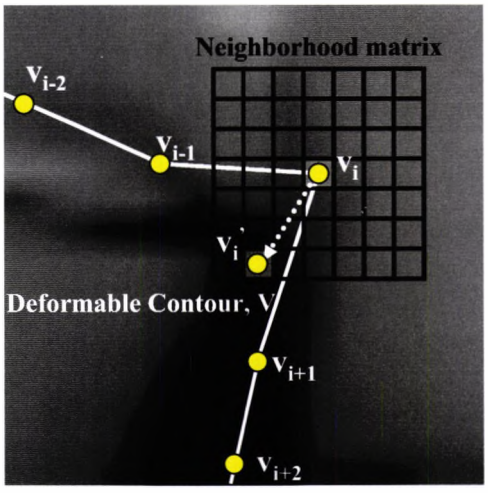


Figure 3.1: Illustration for the deformation of an active contour model.

Finally the snaxel v_i moves towards the new position having minimum energy as making the contour V a circle. The behavior caused by the continuity energy is illustrated in Figure 3.2. Normalisation is required to make the continuity force $E_{continuity}$ independent of the size, location, and orientation of the contour. The normalisation factor $\|\mathbf{V}\|$ is the average distance between the vertices in $\mathbf{V}$:

$$\|V\| = \frac{1}{n}\sum_{i=1}^{n}\|v_{i+1} - v_i\|^2 \tag{3.8}$$

Balloon Force In this work, a balloon force is defined to force the contour to move *inwards* in the absence of external influences [11], i.e. the active contour initialised outside of the target object shrinks under the influence of a reversed balloon force until it approaches the object boundary.

The balloon energy $E_{balloon}$ for a snaxel v_i is expressed as a dot product:

$$e_{jk}(\mathbf{v}_i) = \mathbf{n}_i \bullet (\mathbf{v}_i - n_{jk}(\mathbf{v}_i)) \tag{3.9}$$

where $\mathbf{n}_i$ is the inward unit normal vector at the snaxel v_i. Hence, the balloon energy is smallest at the points farthest from v_i in the direction of $\mathbf{n}_i$. The concept of the balloon force is demonstrated in Figure 3.3. In order to get the normal vector $\mathbf{n}_i$, a tangent vector $\mathbf{t}_i$ at the snaxel v_i is calculated and rotated by 90°. The tangent vector $\mathbf{t}_i$ is defined as:

$$\mathbf{t}_i = \frac{v_i - v_{i-1}}{\|v_i - v_{i-1}\|} + \frac{v_{i+1} - v_i}{\|v_{i+1} - v_i\|} \tag{3.10}$$

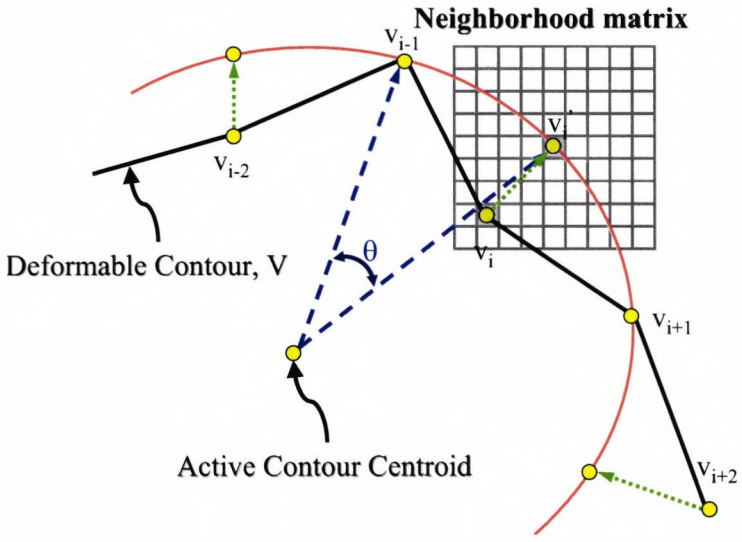


Figure 3.2: Illustration for the movement of a snaxel with respect to the continuity energy: the snaxel v_i moves towards the snaxel candidate v'_i as the continuity energy is minimised.

External Energy

The external energy function attracts the deformable contour to interesting features, such as object boundaries in an image. Any energy expression that accomplishes this attraction can be considered as the external energy. In this work, the *Laplacian* in a 3×3 region is used to provide the gradient information.

The external energy function is expressed as

$$\beta E_{ext}(v_i) = g \cdot \nabla^2 f(v_i) \tag{3.11}$$

where the Laplacian of a 2D function $f(\cdot)$ is a second-order derivative defined as

$$\nabla^2 f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2}. \tag{3.12}$$

The constant g is provided to adjust the relative weights of the terms.

3.3.3 Dynamic programming

The concept of an active contour model using dynamic programming, originally introduced by Geiger [29], is employed in this work. For local optimisation methods such as the greedy algorithm, the optimisation process of the snake's energy function takes place at each snaxel *locally* without regard to

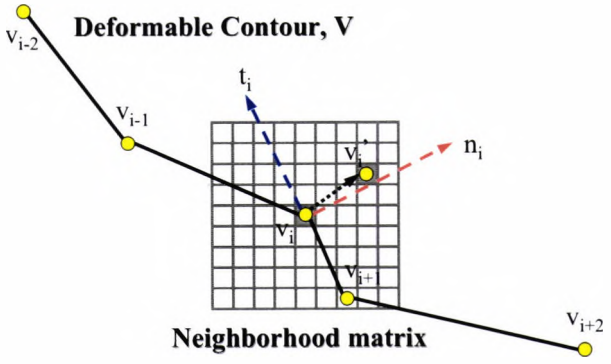


Figure 3.3: Illustration of the snaxel movement according to the balloon force: the position v_i' having minimum energy is chosen as a new candidate for the snaxel v_i.

how the current decision affects the total energy of the solution. However, the implementation of the dynamic programming results in repositioning the snaxels *optimally* within the search neighbourhood for each iteration by considering all possible choices.

The dynamic programming formalism also allows enforcing hard constraints such as a limit on the distance between snaxels on the contour. Finally, in the discrete dynamic programming formulation, the active contour is guaranteed to converge to a final solution in a finite number of iterations since the energy measure is monotonically decreasing with time [1]. This is an important feature for implementing a real-time system since the maximum processing time is always predictable.

The energy minimising function E_{min} consists of a set of sub-functions of which each sub-function corresponds to a pair of adjacent snaxels [29]:

$$E(v_0, v_1, \ldots, v_{n-1}) = E_0(v_0, v_1) + E_1(v_1, v_2) + \cdots + E_{n-1}(v_{n-1}, v_0) \quad (3.13)$$

Each snaxel takes one of m candidates as its updated position while the number of candidates m generally corresponds to the number of the possible locations within the given *neighbourhood matrix*[2]. An approach to solve this minimisation task is to employ *exhaustive searching* by considering the problem as a $m \times n$-dimensional *travelling salesman problem*. However, this exhaustive search dramatically increases the processing time. A more efficient strategy is to use discrete dynamic programming, assuming an updated snaxel position v_i' as a state variable in the i-th decision stage. Dynamic programming determines the minimum not by means of derivatives, but rather by a straightforward search technique. Since the minimisation of the

[2] A 3×3 neighbourhood matrix is used for this application.

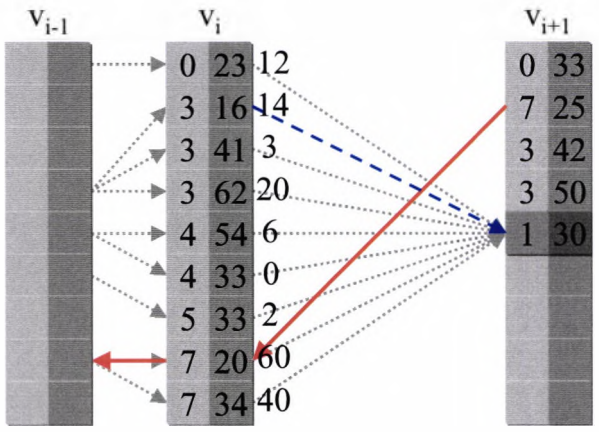


Figure 3.4: Demonstration for a snake using dynamic programming: the blue arrow stands for a link which creates minimal energy, and the red ones represent the traces of an optimal contour.

energy function can be viewed as a discrete multistage search process, the technique of dynamic programming can be applied to the active contour models. Starting from the initial point on the contour, the minimisation problem is treated as one where each of a finite set of minimisation stages corresponds to snaxel positions.

The dynamic programming solution involves generating the sequence of optimal value functions s_i, where for obtaining each s_i a minimisation is performed over a single dimension over v_i. The function s_i is defined as:

$$s_i(v_i) = \min_{v_{i-1}} \left(s_{i-1}(v_{i-1}) + E_i(v_{i-1}, v_i) \right) \tag{3.14}$$

Figure 3.4 shows an example in case of nine candidates per stage, of which each candidate corresponds to an entry in the 3×3 grid of the neighbourhood matrix centred at the current snaxel position v_i. The first column of v_i is a position matrix where each entry represents the index of the candidate chosen from the v_{i-1} neighbourhoods, while the values in the second column show the energy costs propagated from the selected candidates in the first column. The numbers associated with given arrows correspond to the internal energies of the possible choices in decision sets. For each candidate of the next snaxel v_{i+1}, the candidate at the v_i having the minimum sum of the forward cost and internal energy is selected. By tracing back the minimum energy snaxels in the position matrix from the last snaxel column, an optimal contour can be found.

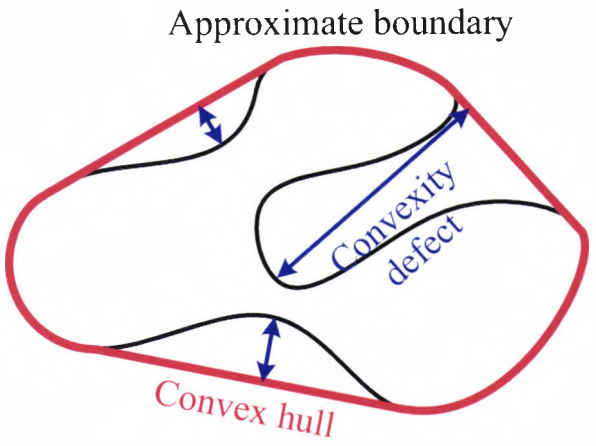


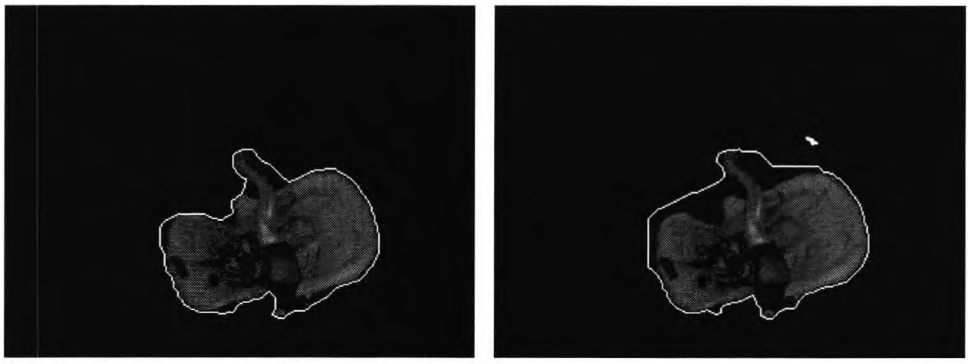
Figure 3.5: The definition of convexity defects with respect to the convex hull created around an object boundary.

3.3.4 Convexity defects driven active contour models

A drawback of using an active contour model is that their convexity properties are poorly understood. Specifically, it has been known that active contour models are non-convex and that solutions are rather complex for real-time applications [106, 17, 18]. Some known solutions often involve discontinuities in the final contour, and this accounts for the phenomenon of convergence to the wrong result.

Since the morphological operation should provide an initial contour *enclosing* a target object, an active contour model shrinks inwards as the energy minimisation proceeds. The outline of a non-rigid object usually involves concave curvatures. Although the *negative* balloon force enforces the snaxels to move towards the centroid of the contour, the contour may not reach the global minima due to the continuity energy which controls the regularity of the active contour. The continuity energy involuntarily hinders the snaxels from approaching to the concave curvatures of the target boundary by decreasing the edge-sensitivity of the active contour. Therefore, the aim of this section is to find a relatively simple solution which improves the segmentation performance at concave boundaries. By analysing the concavity of the target boundary, a set of weighting constants which facilitate the contour deformation to overcome the undesirable effects of the continuity energy is proposed.

A solution to this problem could be to individualise the snaxels according to their geometrical relationships. In the beginning of each active contour evolution, a *convex hull* is created around the initial boundary of the active contour. A sequence of snaxels in the initial boundary exists normally



(a) (b)

Figure 3.6: Typical examples for an active contour model with support of convexity defects: segmentation results (a) with and (b) without concavity analysis. The results are superimposed by the ShadowFlash result.

between two consecutive convex hull vertices while each pair of the vertices form a line segment. For each sequence, a *convexity defect* is defined as the maximum vertical distance between the sequence and corresponding line segment. For example, the convexity defect of a sequence adjacent to/overlapping the corresponding line segment is zero. Finally, the energy function for a given active contour V in Equation 3.4 can be improved as

$$E(V) = \sum_{i=0}^{n} \mu_i \cdot (\alpha \cdot E_{int}(v_i) + \beta \cdot E_{ext}(v_i)) \tag{3.15}$$

where μ_i is the *convexity defect* corresponding to the i-th snaxel. The definition of convexity defects is shown in Figure 3.5 while a typical example demonstrating the influence of the convexity defects analysis is presented in Figure 3.6.

3.4 Experimental results

This section presents the evaluation of the proposed segmentation approach. Image sequences with the resolution of 324×244 were collected at 30Hz by a 12-bit grey scale HDR camera with supplementary active illuminations introduced in Chapter 2. Over 40 different objects including both rigid and non-rigid shapes are used for capturing the test sequences under varying illumination conditions. While the position of the camera remains stationary, a reference background is obtained. The fixed pattern noise caused by the imager as well as the background image components is eliminated by subtracting the reference background from the sequences. The reference background used in this experiment is shown in Figure 3.7. After suppressing the lens distortion using the pre-calibrated lens parameters,

Figure 3.7: The reference background used in the experiment: due to its low contrast, the brightness level of the reference background is manually improved while a passenger seat is shown in the scene (see Figure 6.2). The static fixed pattern noise presented in the input sequences is suppressed by the background subtraction process.

the DoubleFlash technique is applied to the sequences to generate ambient illumination-independent frames. The real-time ShadowFlash creates shadow-free image sequences without reduction of the frame rate.

Figure 3.8 shows an example of successful segmentation for a non-rigid object belonging to the *Adult* class. Figure 3.8(a)-(c) were illuminated by three active light sources from different directions while Figure 3.8(d) is imaged by only ambient illumination. These four consecutive frames formed a *quadlet* of input images which were used for the real-time ShadowFlash technique. Since 33 milliseconds of time delay (30Hz) exist between the frames, motion of the person is noticeable. The result of the ShadowFlash applied to the input quadlet is shown in Figure 3.8(e) where most of the shadow presented in the input quadlet was successfully removed. Figure 3.8(f) shows the result of the texture similarity measurement described in Equation 3.2, and its binarised version is shown in Figure 3.8(g). Following morphological operations, the image fragments were merged together in Figure 3.8(h), and the approximate boundary for initialising an active contour model in Figure 3.8(i) was generated from the largest image blob in Figure 3.8(h). By investigating the concavity of the active contour with respect to the convex hull in Figure 3.8(j), the object of interest is successfully extracted by the active contour model as shown in Figure 3.8(k).

A typical rigid object in the *FFCS* class is segmented in Figure 3.9. As a result of the morphological operations, a small image fragment is accidentally combined with the target object blob resulting in an artifact on the approximate boundary as shown in Figure 3.9(h) and (i). The convexity defect at this artifact is minimal since the artifact is in contact with the convex hull. Despite of the minimal convexity defect, the influence of the artifact is practically removed after the active contour deformation as shown in Fig-

ure 3.9(k). Another segmentation example for the RFCS class is presented in Figure 3.10. The artifact produced by the morphological operations is also successfully suppressed by the snake deformation. Nevertheless, the boundary approximation at the concavity failed regardless of the improved mobility based on the convexity defect analysis.

The advantage of employing the ShadowFlash technique for the segmentation process is evaluated in Figure 3.11. The image on the left side of Figure 3.11(a) was illuminated by a single light source creating the strong cast shadows on the object surface, while most of shadow was suppressed in the ShadowFlash image on the right side. The textural similarity between the object and background in the shadowed region was significantly increased due to the reduction of the average brightness level, and the initial boundary was generated improperly, causing the snake evolution to fail (shown in the left side of Figure 3.11(b),(c) and (d)). On the other hand, the input image on the right side, in which the irradiance of the surface was homogeneous, did not suffer from the effect of shadows. In the right side of Figure 3.11(d), the segmentation performance was dramatically improved after employing the ShadowFlash compared to the one with a single light source. Two more examples are shown in Figure 3.12 and 3.13 for the evaluation of the ShadowFlash effect with respect to the proposed segmentation process. In both cases, the segmentation results were significantly corrupted due to the uneven distribution of illumination.

An example of a less successful case is presented in Figure 3.14. Although the ShadowFlash method improved the homogeneity of the intensity level, the textural property of the target object in the scene was unexpectedly similar to the one of the background. As a consequence, the high textural similarity led the thresholding process to obtain the undesired result for the initial boundary creation.

3.5 Precis

In this section, the problem of extracting the boundaries of non-rigid objects is addressed. The influence of ambient illumination and cast shadows were minimised by employing the DoubleFlash and ShadowFlash techniques demonstrated in Chapter 2. The similarity measurement between each input frame and reference background produced object candidates after thresholding. Morphological operations were employed to combine image blobs as well as to eliminate noisy image fragments. The largest image blob among the candidates was chosen as a target object, and an approximate boundary was extracted from the image blob for initialising an active contour model. Convexity defects were computed to improve the mobility of the active contour while dynamic programming was employed for the energy minimisation

method.

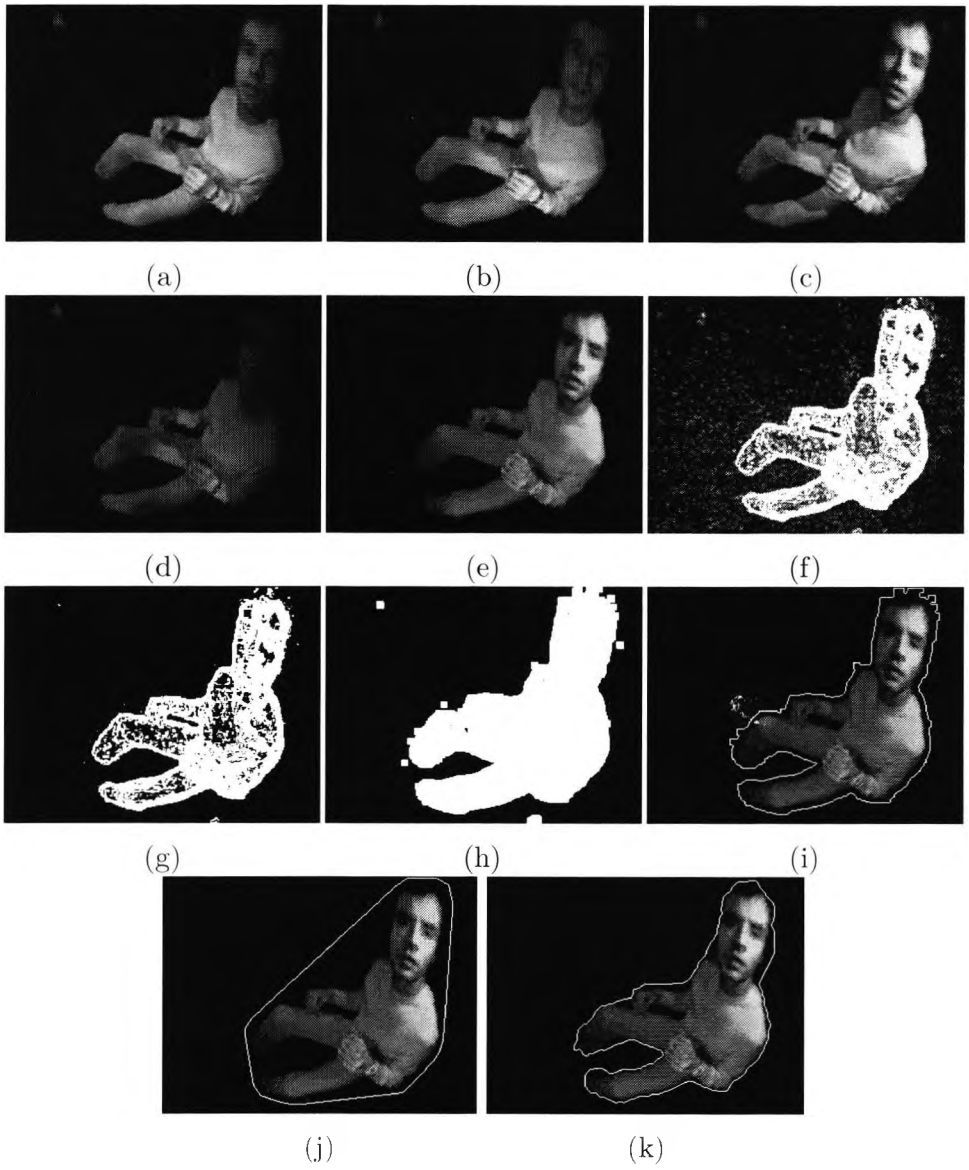


Figure 3.8: Segmentation result applied to the *Adult* class. The background of the input quadlet is beforehand suppressed using the reference background shown in Figure 3.7: (a)-(d) the quadlet of input frames, (f) the texture similarity image, (g) the thresholding result applied to (f), (h) the morphological operation result, (i) the approximate (initial) boundary for the snake evolution, (j) the convex hull and (k) the snake result.

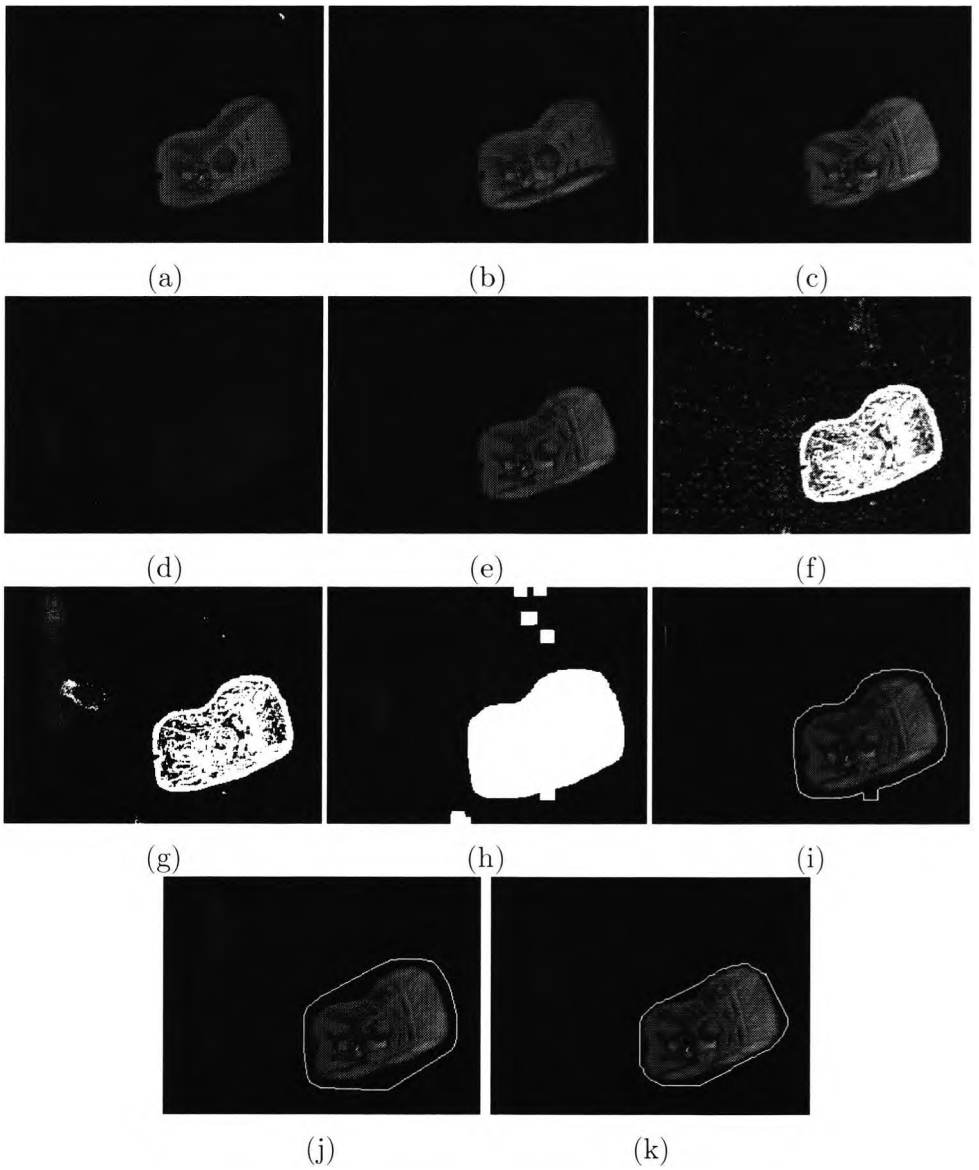


Figure 3.9: Segmentation result applied to the *FFCS* class. The artifact on the initial boundary is successfully removed after the snake evolution: (a)-(d) the input quadlet, (f) the texture similarity image, (g) the thresholding result applied to (f), (h) the morphological operation result, (i) the approximate (initial) boundary, (j) the convex hull and (k) the snake result.

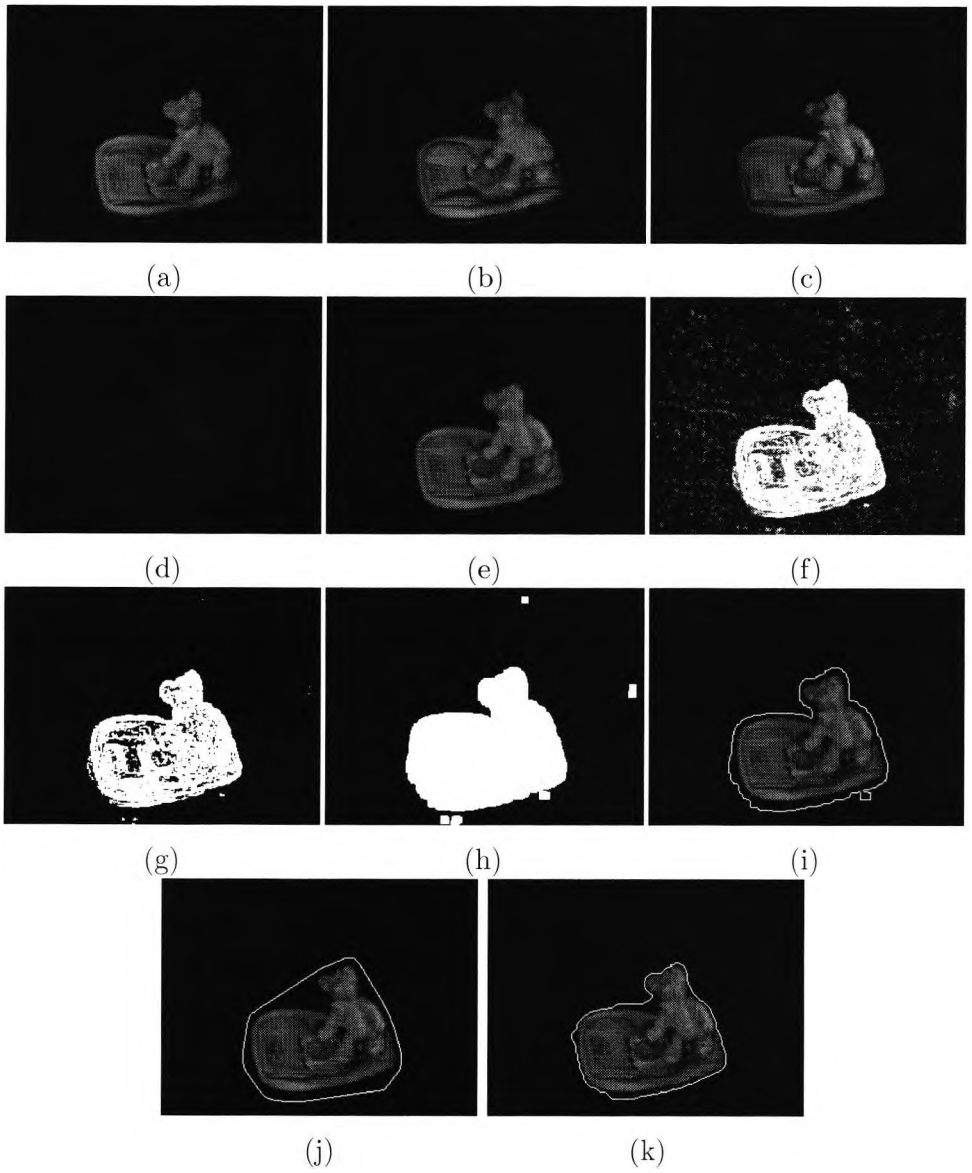


Figure 3.10: Segmentation result applied to the *RFCS* class: (a)-(d) the input quadlet, (f) the texture similarity image, (g) the thresholding result applied to (f), (h) the morphological operation result, (i) the approximate (initial) boundary, (j) the convex hull and (k) the snake result.

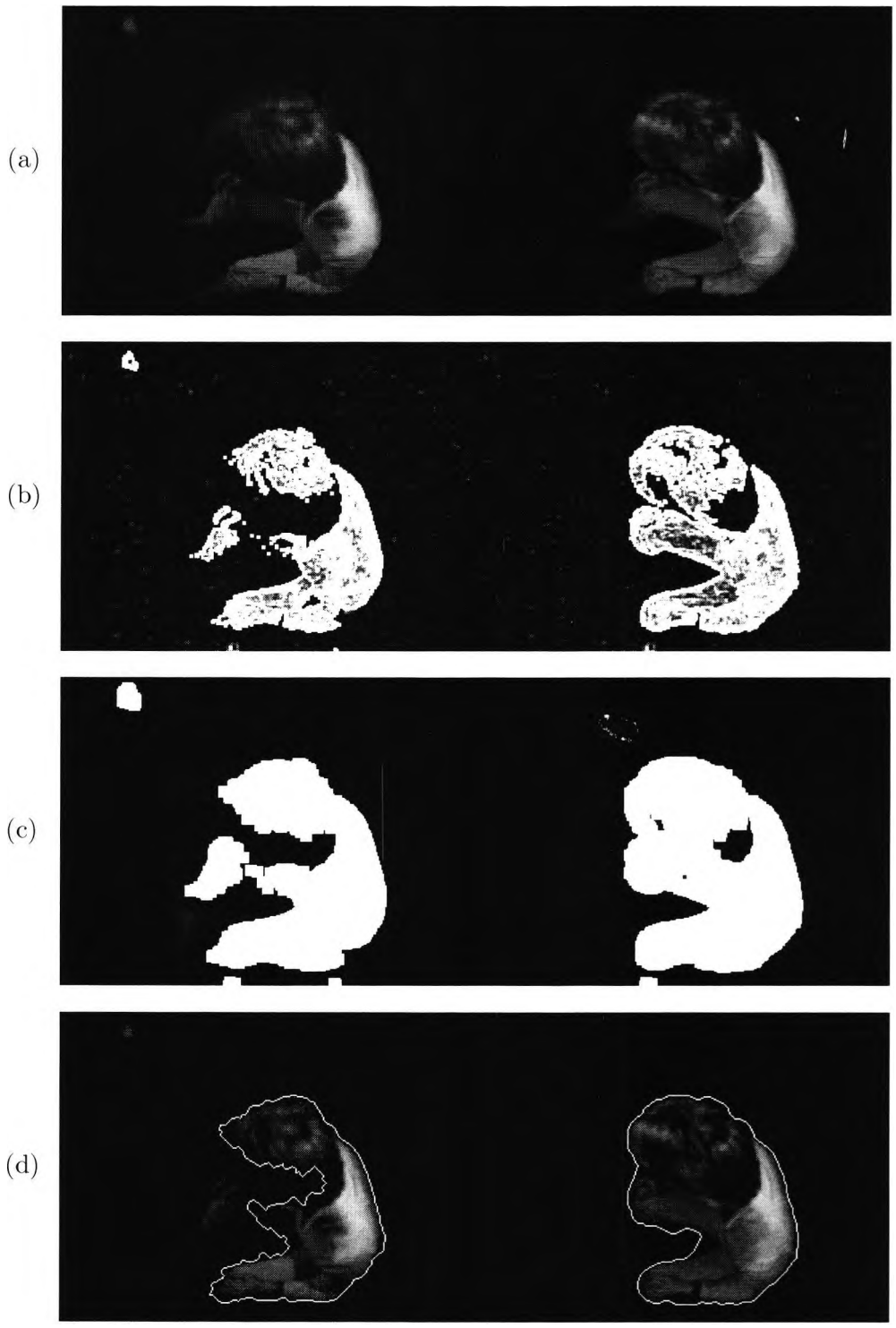


Figure 3.11: The effect of the ShadowFlash technique applied to the segmentation process: (a) *left*: the input image illuminated by a single light source and *right*: the ShadowFlash image, (b) the textural similarity measurement results, (c) the image blobs after conducting the morphological operations, (d) the active contour approximation results.

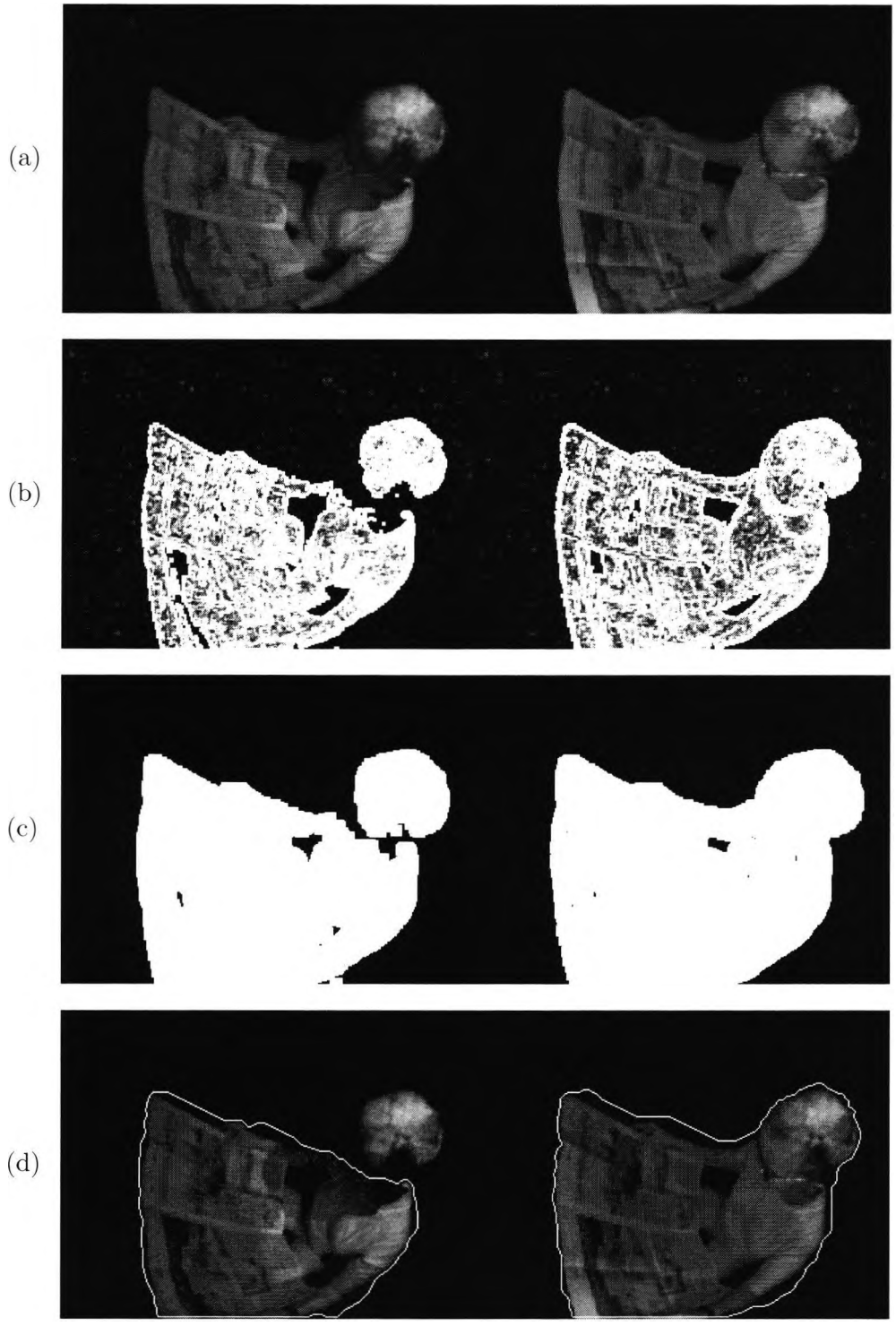


Figure 3.12: Another example for segmentation via the ShadowFlash technique: (a) *left*: the input image illuminated by a single light source and *right*: the ShadowFlash image, (b) the textural similarity measurement results, (c) the image blobs after conducting the morphological operations, (d) the active contour approximation results.

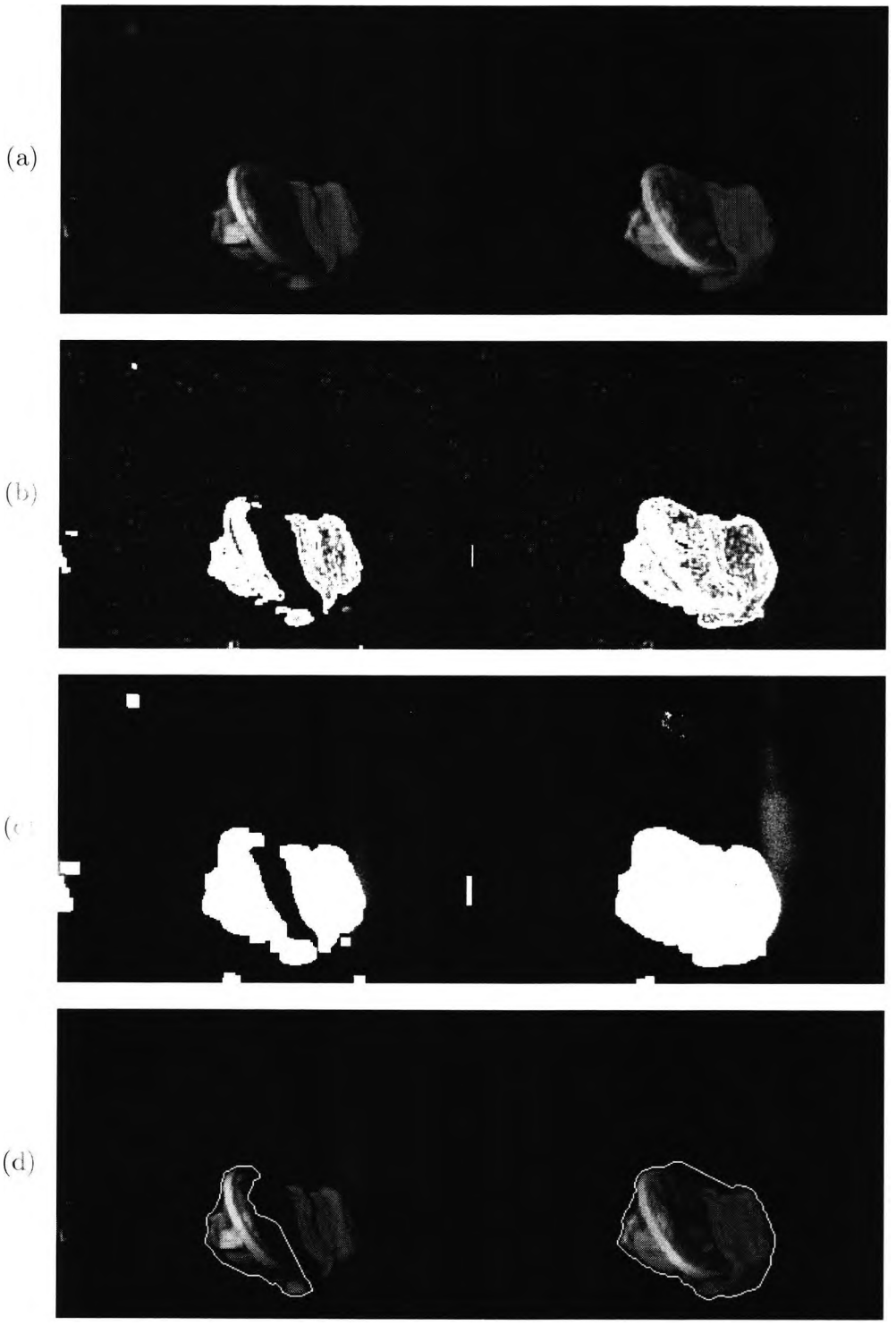


Figure 3.13: The evaluation of segmentation result with the ShadowFlash technique: (a) *left*: the input image illuminated by a single light source and *right*: the ShadowFlash image, (b) the textural similarity measurement results, (c) the image blobs after conducting the morphological operations, (d) the active contour approximation results.

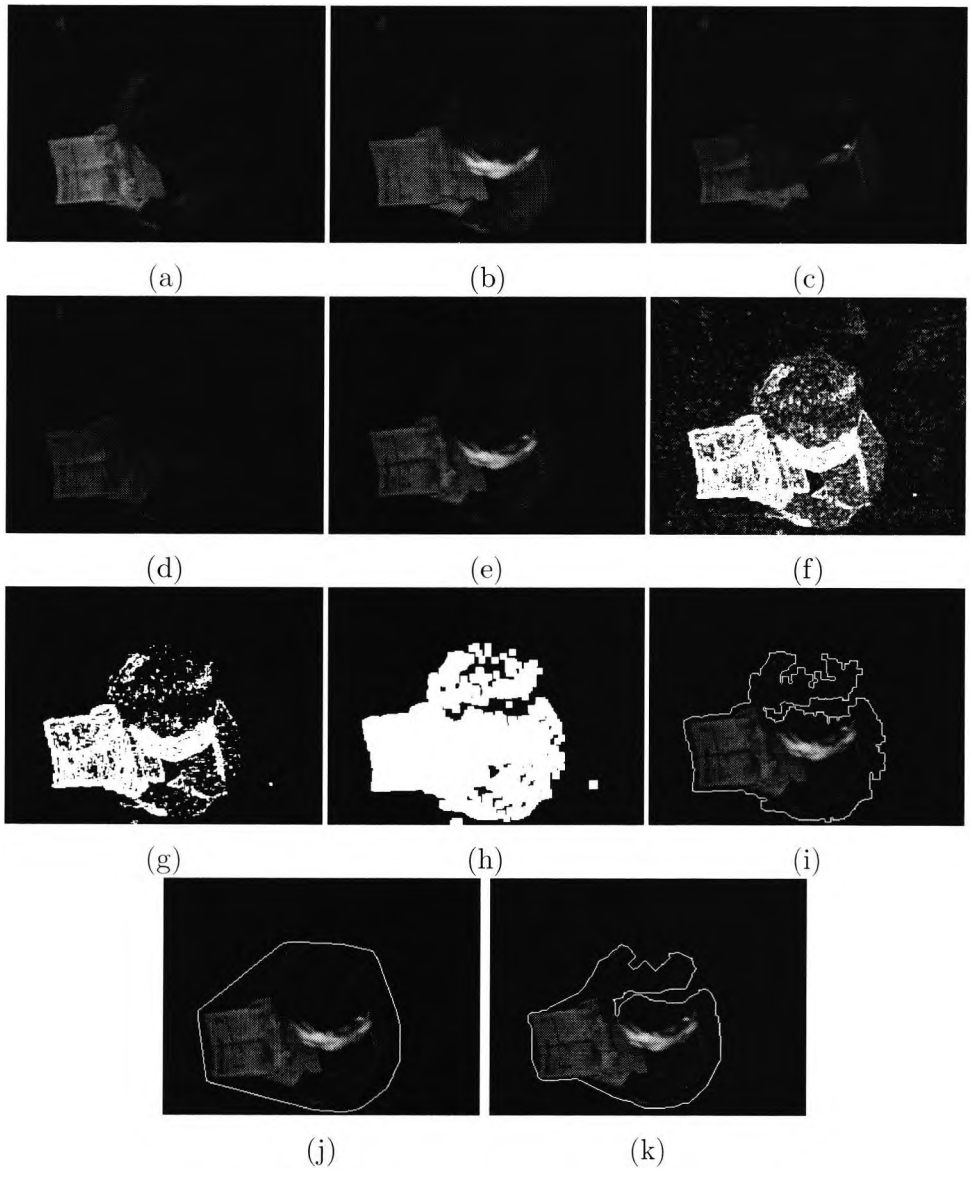


Figure 3.14: Example of less successful segmentation: the boundary approximation failed due to the loss of textural information caused by the high textural similarity between the target object and reference background: (a)-(d) the input quadlet, (f) the texture similarity image, (g) the thresholding result, (h) the morphological operation result, (i) the approximate (initial) boundary, (j) the convex hull and (k) the snake deformation result.

Chapter 4

3D processing: surface reconstruction

4.1 Motivation **74**
4.2 3D sensing techniques and their limitations . . **74**
4.2.1 Ultrasonic imaging 75
4.2.2 Laser scanning 75
4.2.3 Structured lighting 76
4.2.4 Time-of-flight cameras 76
4.2.5 Shape from shading 76
4.2.6 Stereo vision 77
4.3 Stereo vision techniques **78**
4.3.1 Fundamentals 78
4.3.2 Limits and drawbacks 80
4.4 Photometric Stereo Method **83**
4.4.1 Introduction 83
4.4.2 Reflection models 84
4.4.3 Surface normal estimation 85
4.4.4 Surface integration 87
4.4.5 Advantages and drawbacks 89
4.5 Experimental Results **92**
4.5.1 Reconstruction examples 92
4.5.2 Evaluation of the reconstructed surface accuracy . 93
4.5.3 Surface reconstruction of temporal sequences . . . 93
4.6 Precis **94**

4.1 Motivation

Image acquisition always contracts the three-dimensional information of the scene to two-dimensional information of the image due to the projection on the 2D image plane. Therefore the reconstruction of the depth information from the 2D image is a fundamental problem in machine vision. Since fast and non-contact shape measurements are of significant importance in various applications such as industrial inspection, robot vision, surveillance as well as virtual reality, the technologies for three dimensional shape measurements have been in a phase of rapid development for a number of years [42].

For example, faced with the increasing demand for various vision-based vehicle in-cabin monitoring applications, the capability of providing three-dimensional information has become increasingly important. The reconstruction of object surfaces necessitates greater power and bandwidth of image processing hardware to handle the accumulated data as well as special sensors for acquiring images. The higher processing power and bandwidth results in dramatically higher overall system costs. Therefore, the research for implementing a 3D-based vision system with an embedded platform of low cost is of great importance for industrial applications in terms of mass production.

In this chapter, the problem of developing a low-cost single camera solution capable of 3D surface reconstruction is addressed. Various 3D sensing techniques are introduced followed by a discussion on the drawbacks of classical stereo vision techniques. The necessity of a *low-cost* real-time 3D shape reconstruction system has led to the use of photometric stereo methods in which an object surface is computed by integrating shading information. Upon examination of the fundamentals of the photometric stereo method, Wei's algorithm [104] is chosen as the surface integration method of the proposed system. Finally, the advantages and drawbacks of the proposed system compared to the stereo vision based systems are discussed, and the experimental results are presented.

4.2 3D sensing techniques and their limitations

Given tremendous advances in computer vision, it is no longer a problem to process the data from 3D object surfaces even in real-time. The problem which remains is the fast and precise acquisition of the depth information within a large volume and a natural environment. There are various methods for 3D sensing techniques, which deliver the three-dimensional shape and physical dimensions of an object. From the knowledge of the underlying physical and theoretical principles which define the limitations of the shape reconstruction performance, an optimal 3D sensing technique

can be selected to satisfy the environmental requirements of the given application.

The vast number of known 3D imaging techniques are based on three different principles [42]: *triangulation*, *time-of-flight measurement* and *interferometry*. *Triangulation* is the most widely used technique for optical shape measurements. Systems based on this method use mechanically scanned illumination, structured light projection or stereoscopy with several stationary cameras. The rapid progress of optical triangulation, especially (1) active methods with structured lighting, (2) passive methods with digital photogrammetry, and (3) combinations of both, is already a big step towards the goal of real-time stereo vision. *Time-of-flight* techniques detect distance by measuring the time of flight of the envelope of a modulated optical signal. Methods based on *interferometry* measure depth also by means of the time of flight, but they require coherent mixing and correlation of the wavefront reflected from the target object with a reference wavefront.

This section introduces a number of 3D sensing techniques employed to the wide range of industrial applications requiring shape reconstruction. The discussion is focused mainly on, but not restricted to, their suitability to the occupant detection system. Since the aim of this work is to propose an alternative framework with the comparable performance to stereo vision based occupant detection systems, a further discussion is presented on stereo vision techniques in terms of both economical and theoretical aspects in Section 4.3.

4.2.1 Ultrasonic imaging

Ultrasonography is a cheap, widely available and non-hazardous imaging modality for estimation of volumes by analysing 3D ultrasound data which promises accuracy and precision. This technique is especially useful to reliably calculate volumes of organs in medical imaging. However, the attenuation of the ultrasonic beam sets the practical limitation on the range of depth measurement and this restricts the applicability of the ultrasonic imaging technique. Additional difficulties with ultrasound are the significant sensitivity of the propagation speed of sound to pressure and temperature.

4.2.2 Laser scanning

Laser scanning is another non-contact 3D sensing technique based on the triangulation principle where the distance to the object is computed by means of a directional light source. This technology uses the same principle as laser targeting systems in military aircraft for air or ground targeting. By painting a surface using a laser beam, the laser triangulation sensor determines the depth and in some cases, the orientation of the surface being observed.

The more sophisticated versions feature scanning lasers that project a plane instead of a spot onto the surface of the object. The laser plane projection and its degree and direction of distortion can be analysed to render orientation information about the target surface. A major advantage compared to other sensing techniques is the possibility for parameter optimisation for every measure point and which provides high depth resolution and accuracy compared to other depth sensing techniques. The use of laser also means that the sensor is impervious to all ambient lighting changes. However, the laser scanning technique requires complex signal processing to detect quantitative depth information with reasonably high resolution at video frame rate. Eye safety is another important factor to be concerned with prior to employing this technique.

4.2.3 Structured lighting

Structured lighting is an *active triangulation method* which calculates the three-dimensional shape of the object based on the deformation of the light patterns projected on the target object's surface. The calculations are simple and fast so that the shape of the scene could easily be extracted, provided that the feature points of the projected pattern are accurately detected. However, in reality, it is difficult to implement an accurate pattern using an infrared light source due to the constant vibration in the vehicle environment. Furthermore, such patterns may not provide enough resolution for object classification.

4.2.4 Time-of-flight cameras

Recently, a *time-of-flight* (TOF) imager which consists of an array of single point distance measurement units measuring the runtime or phases of the emitted light from a supplementary light source, is of great interest in the industry. The TOF imager has a great advantage in that it directly measures the absolute depth as well as the local brightness and determines a complete distance map of the scene without any delay. Due to the continuing advances of solid-state technology, measurement precision will soon be in the millimeter range, and such cameras will be miniaturised to a size not much larger than the conventional CCD cameras. Nevertheless, as the measurement range is limited by the maximum radiant power, the possibility of violating the eye safety still remains to be solved.

4.2.5 Shape from shading

The reconstruction of non-planar surfaces from a single irradiance image is one of the classical tasks in scene analysis [48]. The shape-from-shading techniques deliver 3D information as normal vectors of the surface elements

from the image irradiance and the known position of the camera and the light sources. By integrating the surface normal vectors, the 3D shape of a target object can be computed. The major advantage of this method is the economical hardware requirements compared to the other 3D sensing techniques. However, as there is insufficient information contained in an arbitrary irradiance image to reconstruct the object surface unambiguously, a shape-from-shading based surface reconstruction system is far from being a complete and video-rate depth image acquisition at the moment.

The problems with shape-from-shading has led to the proposal of the photometric stereo method which reconstructs the object surface by utilising multiple irradiance images [109]. The difficulties with shape-from-shading could be mitigated by acquiring multiple images of the object under different illuminations. Each image provides one constraint on the normal, and therefore two images are sufficient to recover the normal up to small number of possible solutions, and three images yields a unique solution for each image pixel. Photometric stereo enables relaxing the strong smoothness conditions imposed by classical shape-from-shading approaches, and therefore yields more reliable shape estimates.

4.2.6 Stereo vision

The technical realisation of the passive triangulation method called *stereo vision* is the classical approach towards shape recovery which has been used in *photogrammetry* for many decades. Static stereo analysis denotes a very active field of research in computer vision where it is assumed that at least two cameras capture a scene at the same time or within certain time interval. For depth information, multiple cameras are required with known relative positions or self-calibrating methods. Stereo vision is the most widely employed shape reconstruction technique for industrial applications due to the well-established underlying principles as a result of intensive research for decades as well as the resemblance to the human vision system.

Most of currently available commercial occupant detection systems are based on the binocular-stereo vision technique with support of active illumination. However, high computing power with multiple cameras significantly increases the overall system implementation cost. Major advantages of this technology are its independence from object dimensions and maturity of underlying principles. Its drawbacks include more expensive specialised sensors, larger physical size, and more difficult calibration. Further detailed discussion about the stereo vision techniques is presented in the following section.

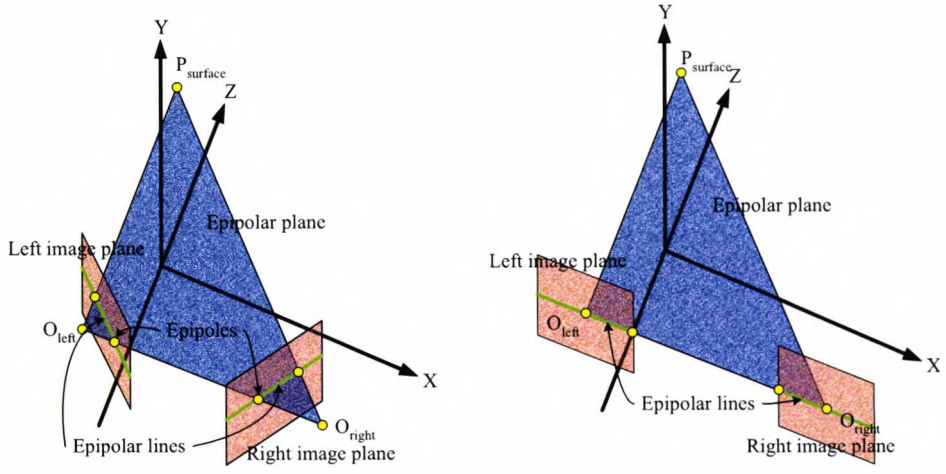


Figure 4.1: Non-standard stereo geometry vs. standard stereo geometry: (a) a general geometric situation with arbitrarily placed cameras and (b) a standard stereo geometry where two image planes are coplanar.

4.3 Stereo vision techniques

This section discusses the drawbacks of the classic multi-camera based stereo vision techniques in terms of the cost-performance effectiveness for industrial applications.

4.3.1 Fundamentals

Figure 4.1(a) shows a general geometric situation with arbitrarily placed cameras. An *epipolar line* is the intersection of an *epipolar plane* with an image plane where the *epipolar plane* is defined by the surface point $P_{surface}$ and the optical centres of the two cameras O_{left} and O_{right}. This epipolar line significantly simplifies the image matching process by constraining all the object points on the epipolar line in the left image plane to be projected into the corresponding points on the epipolar line in the right image plane which is uniquely defined by the left epipolar line. However, since the calculation of the epipolar lines involves complicate triangulation processes, most practical stereo vision systems employ a proper arrangement of cameras which leads to a simplification of the epipolar line estimation called *standard stereo geometry* [48]. With the use of the standard stereo geometry, the epipolar lines in the two image planes coincide with the horizontal scanning rows of the images, and this simplifies the estimation of epipolar lines underlying the binocular image acquisition situation as shown in Figure 4.1(b).

Figure 4.2 illustrates the fundamental concept of the depth estimation in

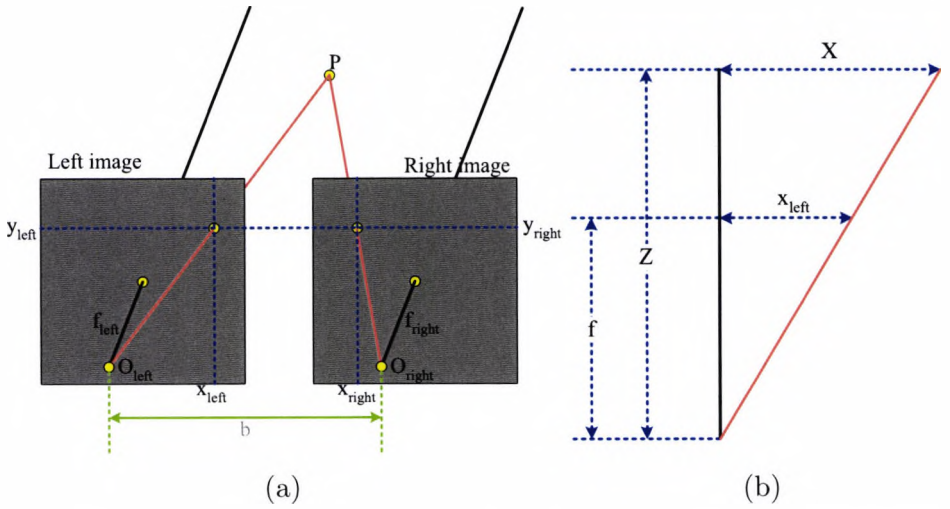


Figure 4.2: Illustration of the basic concept of the stereo vision technique: (a) the standard stereo geometry where two image planes are coplanar and (b) the triangulation process in case of the standard stereo geometry. **P** denotes a point on the object surface while $\mathbf{O}_{left}$ and $\mathbf{O}_{right}$ represent the camera origins. Two black lines stretched from the origins are the optical axes of two cameras. The focal length f is defined as the distance between the camera origin and the image centre.

case of the standard stereo geometry. The a pair of image planes generated by two different cameras are coplanar while the focal lengths of the cameras f_{left} and f_{right} are identical. In case that the coordinate system is oriented at the left camera, the focal points of the left and right camera lie at the camera origins $\mathbf{O}_{left} = (0, 0, 0)$ and $\mathbf{O}_{right} = (b, 0, 0)$, respectively. While the distance between two cameras defines the camera baseline b, the two optical axes of the cameras are assumed to be parallel. Suppose that a object surface point $\mathbf{P} = (X, Y, Z)$ is projected into two *corresponding image points*

$$\mathbf{p}_{left} = (x_{left}, y_{left}) \quad \text{and} \quad \mathbf{p}_{right} = (x_{right}, y_{right})$$

in the left and right image plane, respectively. The *disparity* between the projected image points with respect to the projected image point $\mathbf{p}_{left}$ is defined as

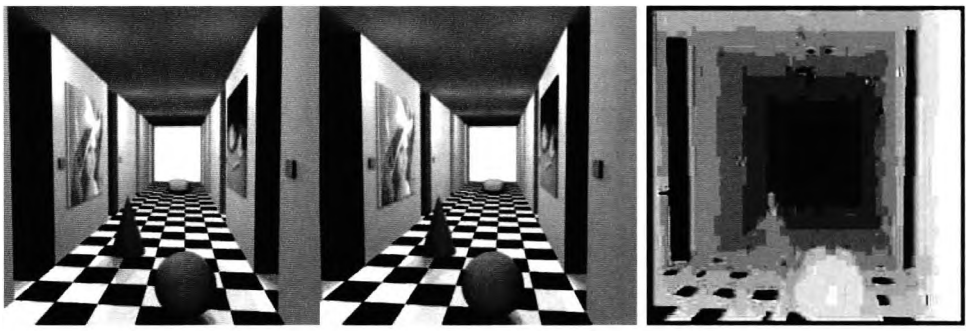
$$\Delta(x_{left}, y_{left}) = (x_{left} - x_{right}, y_{left} - y_{right}).$$

Since the two image planes are coplanar, y_{left} is identical to y_{right} and the definition of the disparity can be simplified to

$$\Delta(x_{left}) = (x_{left} - x_{right}).$$

The triangulation process for estimating the distance between the object surface and left camera is illustrated in Figure 4.2(b). The depth Z with respect to the focal points of the two cameras is defined as

$$Z = \frac{f \cdot X}{x_{left}} = \frac{f \cdot (X - b)}{x_{right}}$$



(a) (b) (c)

Figure 4.3: The result of a poor correspondence analysis due to the high textural similarity caused by the lack of sufficient textures at particular regions: (a) and (b) a synthetic stereo image pair, and (c) the disparity map produced by a region-based correspondence matching algorithm.

where $f = f_{left} = f_{right}$. By eliminating X from the equations, the depth Z is finally expressed as

$$Z = \frac{b \cdot f}{x_{left} - x_{right}} \tag{4.1}$$

which is a function of the baseline b, the focal length f and the disparity between two corresponding image points. Since the focal length and baseline are constant, the absolute depth of an object is proportional to the distance between the corresponding pixels.

4.3.2 Limits and drawbacks

Robustness to homogeneous textures

Stereo vision algorithms operate by *correspondence analysis* which locates same features in both images. Using the geometrical relationship between the cameras and the location of the features in each image, the depth of each feature can be triangulated and finally used for constructing a depth map. The challenge is the successful identification and location of these corresponding object features in both camera images.

There have been a vast number of research addressing the computation of the disparity between two corresponding points in stereo image pairs for decades. These research are mostly based on only two different approaches: *feature-based* and *intensity-based* correspondence analyses.

If a correspondence matching is based on comparing image features such as edges, corners, etc., then this technique is called *feature-based correspondence analysis*. This approach is less sensitive to the illumination variations. Furthermore, the ambiguities in correspondence analysis are significantly reduced compared to intensity-based approaches since the number of possible

candidates for the matching is considerably smaller. Therefore, the accuracy of determined disparities are usually higher than the intensity-based approaches. Since some problems of the feature-based stereo vision system are immediately apparent, the feature-based approaches are *less suitable* to the vision applications where the shapes of target objects are deformable (non-rigid) such as the occupant detection systems. First, the depth map is sparse since edge features are required in both images to produce points of correspondence. Second, the technique fails to extract depth data at points where the boundary feature aligns with line separating the camera geometry. Because of these limitations, there is currently some on going work on combining stereo viewing with other techniques, notably conventional 2D image segmentation and shape from shading and texture methods [66].

Intensity-based correspondence analysis is based on the assumption that corresponding pixels have a similar intensity value. Since identical intensity values can occur in many points of a given image, a set of neighbouring pixels in an image window are used for the correspondence analysis by employing some *similarity measuring* functions. Assuming that the illumination conditions of a target scene are controllable, the performance of this approach could become comparable to that of the feature-based techniques. However, areas with no significant texture or with repetitive texture like a chess table increase the image matching ambiguity and mostly lead the system to fail to deduce the disparity accurately. For example, since the clothing, hair, and skin of passengers in a vehicle often do not include significant textures, the occupant detection systems based on the intensity based stereo vision frequently fail to provide a dense depth map of the target occupant. Figure 4.3 shows the result of correspondence analysis performed on a pair of synthetic stereo images by a region-based stereo algorithm. The disparities on the checker-textured floor are poorly estimated due to the insufficient similarity information caused by the absence of textures in these areas.

Implementation cost and hardware complexity

There are several fundamental factors which affect the implementation cost of practical stereo vision systems: (1) Although the market prices of imagers have been dramatically lowered due to increasing demand of such digital imagers, the necessity for *multiple imagers* is the major factor to increase the hardware costs. (2) To minimise the influence of potential illumination changes in a target scene, it is inevitable to employ *active illuminations* for most of the stereo vision systems exposed to natural illumination conditions. (3) To improve the accuracy of the correspondence analysis, stereo vision systems tend to employ *high image resolution* as well as *subpixel* techniques. The higher image resolution also adds to the overall system cost due to the necessity of faster computation and wider bandwidth of image processing hardware as well as complex electric wiring. (4) Since a certain amount

of the baseline distance between two cameras must be guaranteed for the satisfactory depth resolution, the dimension of a single stereo vision system becomes significantly large. Assuming that several passenger seats have to be analysed independently, the integration of *multiple* stereo vision-based occupant detection systems will not be possible within the vehicle interior roof of the limited space.

Sensitivity to mechanical vibrations

For example, an occupant detection system is in most instances exposed to the wide range of mechanical vibrations. This greatly detracts from the stability of the positioning of the sensor arrays during image acquisition, and the system should be repaired or re-calibrated after the use of several years. This problem may be solved by using not only a single image row for the correspondence analysis, rather a certain interval of rows. However, this solution still can not entirely overcome the problem of constant vibrations and/or unintended strong mechanical impacts. In case that both stereo cameras are sufficiently well mounted, the problem with vibrations can be reduced significantly. Nevertheless, this hardware requirement would decrease the applicability of such stereo vision systems.

Depth resolution

The geometrical laws underlying the stereo vision techniques also restrict the feasibility of the stereo vision based applications in practice. The *depth resolution* is defined as the number of the *depth steps* used for describing the reconstructed 3D surface of an object. Since the focal length as well as the baseline distance of a stereo vision system is constant over time, the depth is only a function of the *disparity* between the corresponding pixels in the stereo image pair as shown in Equation 4.1. Finally, the *depth resolution* can be derived by computing the difference between the *minimum* and *maximum* disparities within the estimated depth range of the target object. If the spatial resolution of the camera is not sufficiently high, the system may have difficulties to identify the target object due to the limited number of the estimated depth steps in case that only a limited physical distance between the imager and target objects is allowed.

For occupant detection systems, the favourite location for the cameras is the electronic console box near to the room mirror. Assuming that the behaviour of a passenger follows a general pattern, the distance between the passenger and cameras should lie between 50 to 100 centimetres. Suppose that the focal length is 2.5 millimetres, the pixel size of the camera is 7.5 micrometres, and the baseline distance between two identical cameras is 5 centimetres. According to Equation 4.1, the disparity dx_{far} at 100 centime-

tres away from the cameras is calculated as

$$\begin{aligned} 1 &= \frac{0.05 \times 0.0025}{7.5 \times 10^{-6} \times dx_{far}} \\ dx_{far} &= 16.7 \text{ pixels.} \end{aligned}$$

Similarly, the disparity dx_{near} at 50 centimetres is

$$\begin{aligned} 0.5 &= \frac{0.05 \times 0.0025}{7.5 \times 10^{-6} \times dx_{near}} \\ dx_{near} &= 33.3 \text{ pixels.} \end{aligned}$$

Finally, the *depth resolution* Δdx of an object located between 50 and 100 centimetres is computed as

$$\Delta dx = dx_{near} - dx_{far} = 16.7 \text{ pixels.}$$

The result implies that the depth map reconstructed by the stereo vision without support of the *subpixel* resolution has only *17 depth steps* in the vehicle environments, and this may not deliver sufficient information for the classification.

4.4 Photometric Stereo Method

4.4.1 Introduction

The *photometric stereo method* (PSM) is an extended version of the *shape-from-shading* (SFS) using multiple light sources, which constructs the *relative depth* of the object by using its reflection properties. Unlike the shape-from-shading, which suffers from the lack of sufficient information in an arbitrary irradiance image to reconstruct the object surface unambiguously, it was successfully proven that the photometric stereo method performs the surface recovery with greater ease, especially when there are more than three light sources.

Since multiple illuminations are already employed for the ShadowFlash method, it is simple to apply the photometric stereo method, for there is no need to provide additional hardware for such an implementation. The problem of using the photometric stereo method for this application is that any abrupt movements of objects in-between two successive frames may cause a significant distortion of the recovered surface. However, after months of repeated tests, in reality it has been concluded that the amount of distortion caused by motion is acceptable for applications which do not need to make decisions frame-wise, especially for systems which do not require high spatial resolution of the scene. The frame rate of the imager is also a primary factor which influences the reconstruction performance.

The overall task of the photometric stereo method involves two major procedures: estimation of surface normals, and integration of the object surface from the normal vectors. The estimation of the surface normal vector could be performed independent of *albedo* by solving irradiance equations supported by *a priori* information about the direction and power of the illumination sources [48]. The Frankot-Chellappa algorithm [26], based on minimising integrability conditions in the frequency domain, is employed after a minor modification to improve its robustness for small artifacts caused by motion regardless of its disadvantage in the processing time.

4.4.2 Reflection models

The amount of light encoded into the gray value of a particular pixel of a digital image can be seen as the result of interactions between surface materials and light sources [48]. Since all the shading-based shape recovery approaches including the photometric stereo method are influenced by the lighting conditions as well as the reflection characteristics of the observed objects, it is necessary to model the properties of both the illumination and object materials.

To simplify the process of the surface orientation estimation, a *Lambertian surface* is assumed in this work. A *Lambertian surface* is a surface of perfectly matte properties, which means that it adheres to *Lambert's cosine law*. *Lambert's cosine law* states that the reflected or transmitted luminous intensity in any direction from an element of a perfectly diffusing surface varies as the cosine of the angle between that direction and the normal vector of the surface. As a consequence, the luminance of that surface is the same regardless of the viewing angle.

The reflection properties can be represented relatively easily by a so-called *reflectance map* originally introduced in [36]. The reflectance map $R(p,q)$ determines the proportion of light reflected as a function of p and q, where the quantity (p,q) is referred to as the *gradient vector* defined as

$$grad(Z) = (p,q)^T = \left(\frac{\partial z}{\partial x}, \frac{\partial z}{\partial y}\right)^T \tag{4.2}$$

at the image point (x,y) with a surface function $z = z(x,y)$.

Finally the *Lambertian reflectance map* can be derived using normalised dot products of the *surface normal vector* $\mathbf{n} = (p,q,-1)^T$ and the *illumination direction vector* $\mathbf{s} = (p_s, q_s, -1)^T$ based on the *radiance equation* [39]:

$$\begin{aligned} R(\mathbf{n}^\circ) &= E_0 \cdot \rho \cdot \mathbf{n}^{\circ T}\mathbf{s}^\circ \\ &= E_0 \cdot \rho \cdot \frac{p \cdot p_s + q \cdot q_s + 1}{\sqrt{p^2+q^2+1} \cdot \sqrt{p_s^2+q_s^2+1}} \\ &\text{where } \mathbf{n}^\circ = \frac{\mathbf{n}}{\|\mathbf{n}\|} \quad \text{and} \quad \mathbf{s}^\circ = \frac{\mathbf{s}}{\|\mathbf{s}\|} \end{aligned} \tag{4.3}$$

where E_0 denotes the light source irradiance and ρ represents the *albedo* which describes the ratio of reflected to incoming radiation.

Assuming that each surface element receives the same irradiance, the scene radiance, and hence image intensity, depends only on the surface normal defined by the surface gradients p and q. Since the viewed image intensity is directly proportional to the surface radiance [32], the image intensity $E(x, y)$ and reflectance map $R(p, q)$ can be made equivalent by setting the proportional constant c to one:

$$E(x, y) = c \cdot R(p, q) \simeq R(p, q). \tag{4.4}$$

This result, called the *image irradiance equation*, is the most important tool to describe the relationship between irradiances, scene radiances, and surface gradients as the equation is the basis of the shading based shape recovery methods [48].

4.4.3 Surface normal estimation

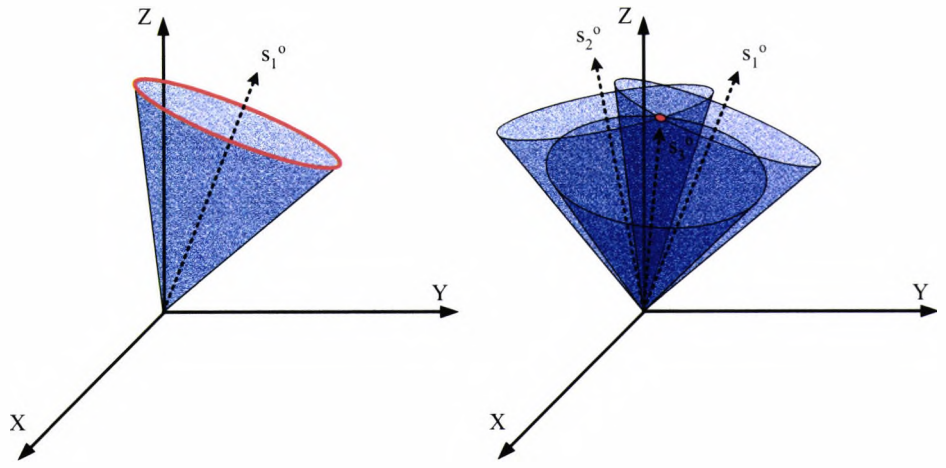
In case that the illumination direction and the reflectance function of a given surface are known, a constraint could be provided to the orientation of the surface normal. Assuming that a Lambertian surface is illuminated by a distant point source of intensity E_0, Equation 4.3 associated with Equation 4.4 can be rearranged as

$$\mathbf{n}^{\circ T}\mathbf{s}^{\circ} = \frac{E(x, y)}{E_0 \cdot \rho} = \text{constant}. \tag{4.5}$$

Since the intensity value of the surface $E(x, y)$ is constant, the passible candidates of the surface orientation can be displayed by a right circular cone as shown in Figure 4.4(a). This explains the difficulty of the surface normal estimation using the *shape-from-shading* approach with a single light source. Since there is an infinite number of possible solutions, the surface orientation can only be determined uniquely in special cases requiring sufficient constraints. To determine local surface orientation without such constraints, additional information is required. The simplest approach is to take *multiple* shaded images illuminated by the light sources in different positions, rather than one. This is called the *photometric stereo method*, originally introduced in [109].

In case of three light sources, the image irradiance equations of the system are defined as

$$\begin{aligned} E_1 &= E_{01} \cdot \rho \cdot \mathbf{n}^{\circ T}\mathbf{s}_1^{\circ} \\ E_2 &= E_{02} \cdot \rho \cdot \mathbf{n}^{\circ T}\mathbf{s}_2^{\circ} \\ E_3 &= E_{03} \cdot \rho \cdot \mathbf{n}^{\circ T}\mathbf{s}_3^{\circ}. \end{aligned}$$



(a) (b)

Figure 4.4: Illustration of the set of solutions of the different shading based shape recovery methods represented by the right circular cones. The *red* area shows the candidates of the surface orientation: (a) the shape-from-shading approach with a single light source and (b) the photometric stereo method with three light sources.

Since the normalised surface normal $\mathbf{n}^\circ$ is a part of all equations, the equations can be represented in a matrix form:

$$\mathbf{E} = \rho \mathbf{E_0} \cdot \mathbf{S} \cdot \mathbf{n}^\circ \tag{4.6}$$

where the image irradiance matrix $\mathbf{E}$ is

$$\mathbf{E} = (E_1, E_2, E_3)^T$$

and the light source irradiances are represented as the diagonal matrix

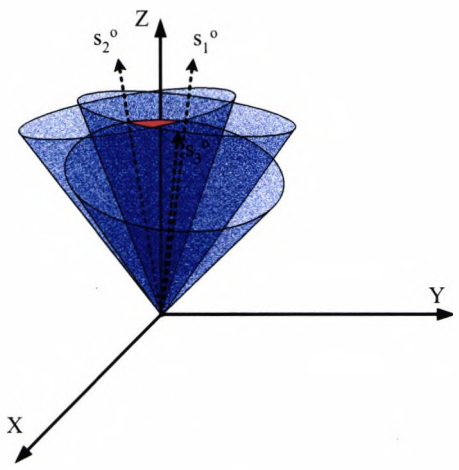
$$\mathbf{E_0} = \begin{pmatrix} E_{01} & 0 & 0 \\ 0 & E_{02} & 0 \\ 0 & 0 & E_{03} \end{pmatrix}.$$

The illumination direction vectors are also described by the matrix

$$\mathbf{S} = (\mathbf{s}_1^\circ, \mathbf{s}_2^\circ, \mathbf{s}_3^\circ)^T = \begin{pmatrix} s_{1x}^\circ & s_{1y}^\circ & s_{1z}^\circ \\ s_{2x}^\circ & s_{2y}^\circ & s_{2z}^\circ \\ s_{3x}^\circ & s_{3y}^\circ & s_{3z}^\circ \end{pmatrix}.$$

After inverting the matrices $\mathbf{E_0}$ and $\mathbf{s}$, the unit surface normal $\mathbf{n}^\circ$ scaled by the albedo ρ can be derived as

$$\rho \cdot \mathbf{n}^\circ = \mathbf{S}^{-1} \cdot \mathbf{E_0}^{-1} \cdot \mathbf{E}. \tag{4.7}$$



(b)

Figure 4.5: Illustration of the estimated surface normal candidates in the existence of various noise sources: the red area represents a set of the possible surface normals with errors.

Since the surface normal vector $\mathbf{n}^\circ$ is a *unit* vector, the equation can be rearranged by dividing the equation by its norm:

$$\mathbf{n}^\circ = \frac{\mathbf{S}^{-1} \cdot \mathbf{E_0}^{-1} \cdot \mathbf{E}}{\|\mathbf{S}^{-1} \cdot \mathbf{E_0}^{-1} \cdot \mathbf{E}\|}. \tag{4.8}$$

Finally the surface normal vector becomes *independent* of the surface reflectivity ρ. This property is helpful for the realisation of an *albedo-independent* photometric stereo method, since only the irradiances and the directions of the light sources have to be known, and discontinuous albedo changes are more likely for real-world objects. Figure 4.4(b) illustrates that the unique solution of the photometric stereo methods using three light sources is found at the intersection of three right circular cones.

4.4.4 Surface integration

Since the surface normals obtained by the photometric stereo methods only describe the orientations of the surface, these information still have to be transformed into depth to provide sufficient three-dimensional information of the object. The estimation process of the surface normals involves various noise sources such as sensor noise, optical lens distortions, non-linear transfer function of the sensor, etc. Therefore, the shape reconstruction accuracy depends significantly on the performance of such a transformation module. Figure 4.5 shows that the number of the possible surface normal solutions under the influence of noise becomes larger compared to Figure 4.4(b). The

objective of these transformation algorithms is to minimise the ambiguities in these solutions.

There are two prominent approaches to recover the object surface from the surface normals: *local propagation* and *global minimisation* [49].

Local propagation

Local propagation approaches start from a single reference surface point or a set of surface points where the shape either is known or can be uniquely determined and propagate the shape information across the whole image. In [15] Coleman proposed a surface reconstruction method which starts the integration in the middle of the gradient field scanning all four quadrants in column direction where their initial path forms a cross in the array. The averaged surface normal is computed from *two points* in sequence, defining a surface tangent from the previous point to the next location. Healey extended this to an *eight-point* method in [34]. Another scanning path parallel to the x- or y-axis, where the gradient values were averaged for obtaining increments in height, was proposed by Wu in [112]. Bichsel [5] developed an efficient minimum downhill approach which directly recovers depth and guarantees a continuous surface. Given initial values at the singular points, the algorithm looks in eight discrete directions in the image and propagates the depth information away from the light source to ensure the proper termination of the process. Although all these algorithms are relatively faster than the minimisation-based techniques, such local propagation algorithms have the drawback that errors may be propagated without any control mechanisms.

Global minimisation

Global minimisation approaches compute the solution which minimises an energy function over the entire image. The function can involve the brightness constraint and other constraints such as the smoothness constraint, the integrability constraint, the gradient constraint, and the unit normal constraint [38, 39, 37, 26, 118, 110, 117, 104].

Frankot and Chellappa presented a solution to enforce strict integrability in an iterative shape-from-shading algorithm while the *integrability* constraint is based on minimising the following function [26]:

$$\iint \left((Z_x - p)^2 + (Z_y - q)^2 \right) dxdy \tag{4.9}$$

where the gradient values on the target surface $Z(x, y)$ are Z_x and Z_y, and p and q are the estimated gradients obtained by shading-based shape reconstruction methods. The original solution for the surface slopes is projected

onto a subspace of surfaces which can be represented by a set of Fourier basis functions, and fulfills automatically the integrability constraint. The advantage of this algorithm over other approaches which incorporate integrability by the use of a penalty function, is the enforcement of strict integrability, whereas the penalty term affects the solution only *close* to integrability.

In [104], Wei introduced another constraint to deal with the local deflection of the surface area and curvature, and the cost function to minimise was extended to

$$\begin{aligned} & \iint \left((Z_x - p)^2 + (Z_y - q)^2 \right) dxdy \\ + \quad & \lambda \iint \left(Z_x^2 + Z_y^2 \right) dxdy \\ + \quad & \mu \iint \left(Z_{xx}^2 + 2Z_{xy}^2 + Z_{yy}^2 \right) dxdy \end{aligned} \tag{4.10}$$

where the subscripts represented the partial derivatives of the surface. The non-negative parameters λ and μ established a trade-off between the constraints. By formulating the special case of Fourier basis functions, the optimisation problem is interpreted into the frequency domain while a computationally efficient implementation was possible by using the fast Fourier transform. This approach is especially useful when the boundary conditions are unknown and the target scene is composed of fairly complicated surfaces.

4.4.5 Advantages and drawbacks

Texture dependency

As discussed in Section 4.3.2, the problem with employing the stereo vision techniques is these techniques' inability to handle surfaces without significant textures. This is not suitable for applications which require high robustness in the case that the operational environments experience a limited range of textural variations. Since the depth estimation of the shading-based surface reconstruction methods does not depend on the accuracy of the correspondence analysis but only depends on the reflection property of the given input images, the *dense* representations of surface shape is guaranteed unless the dynamic range of the scene exceeds the maximum capability of the imager.

Absolute depth vs. relative depth

In contrast to the passive stereo vision approaches, one problem with the shading-based shape recovery methods is that only *relative depth* can be produced rather than *absolute depth*. This means that the prediction of the absolute distance between the imager and object could be a difficult task

for the shading-based approaches. The rough approximation of absolute depth from the given relative depth might still be possible if the geometry of the imaging system is completely known. However, the design of such an intensive geometry calibration process requires sophisticated algorithms with higher computational complexity, which increases the system implementation cost as well as the difficulty of maintenance, and reduces the feasibility of practical low-cost 3D imaging systems. Furthermore, the surface description based on relative depth already provides sufficient amount of geometrical information for applications where only the determination of object types is considered.

Cost and complexity of implementation

The efficient and economical implementation as well as lower maintenance cost are critical factors to assess the feasibility for industrial mass production. The proposed system has the great advantage over the stereo vision-based systems of a reduced number of hardware components. For example, since extremely limited space is reserved for a vehicle interior monitoring system, it has been a great problem for the system developers to assemble the system into one compact package which suits the customer's taste and such a constraint is not easy to overcome for a stereo vision-based system due to its minimum baseline alignment requirement. In contrast, the underlying technique of the proposed system requires only one imager with no geometrical constraints. This makes the package design simpler and straightforward as greater freedom of placement is allowed. Considering most of the stereo vision based systems employ active illuminations for obtaining satisfactory brightness in the field of view, the additional costs of the required light sources are insignificant compared to the stereo vision system. The locations of the illumination sources are chosen under geometrical constraints imposed by the ShadowFlash technique discussed in Section 2.4. However, the problem of wiring the light sources in the limited space in a vehicle roof could become a negative factor to increase the implementation complexity.

Surface distortion

There are various noise sources which can cause distortion of the recovered object surface. In this section, three major distortion factors which often occur in the proposed system are discussed.

Self-reflections and cast shadows Any *intra-object* or *inter-object reflections* can result in the false computation of surface normals due to the *unevenly biased* surface irradiances. Since the prediction of the self-reflections assumes the orientations of all the surfaces affecting the target surface are known beforehand, it is difficult to completely eliminate the influence of the potential reflections. Similar to

the self-reflections, shadows cast by strong light sources is another primary factor to disturb the surface estimation process. Based on the proposed shadow removal technique discussed in 2.4, it is possible to generate shadow-free inputs for the photometric stereo method by simulating *three artificial infinite illumination planes* which illuminate the surface from different directions. For such a simulation, *nine* light sources located at different positions are required. In this case, the advantages of the implementation cost and the system compactness over the stereo vision based systems will be compromised. The inter-frame delay will become another significant factor to reduce the quality of the reconstructed surface.

Imager characteristics Since the photometric stereo method assumes that the surface intensity is directly proportional to the radiance power of the employed light source, the use of a HDR imager could cause surface distortion due to its *non-linear response* characteristics. This problem could be resolved by providing an appropriate reflectance map or by employing a linearisation process similar to *gamma correction*. Any functions affecting the sensitivity characteristics of the imager such as the *automatic gain control* (AGC) must be disabled before the acquisition. Improper design of a lens-undistortion could also cause serious defects on the recovered surface as a consequence of the interpolation.

Motion Since the theory underlying the photometric stereo method is based on the assumption that no motion is present in the field of view during the acquisition process, unrecoverable distortions could occur on the reconstructed surface as a consequence of object's movements. The proposed system can be influenced by the motion caused by the frame delay due to the sequentially captured input images. There are mainly three approaches to minimise the influence of object motion: (1) By assuming *sufficient frame rate* of the imager, the amount of distortion caused by motion would be acceptable for applications which do not require frame-wise decision. In general, low time requirement for image acquisition only allows the errors caused by the fast moving segments, which contain minor descriptive information about the target object. (2) *Downsampling* is another approach used to minimise the ambiguity of deducing the irradiance vector $\mathbf{E}$ by observing the lower spatial frequency components. Specifically, this method is preferable in the case that only the coarse description of the object shape is required for subsequent processes. (3) *Motion detection* techniques such as *optical flow* may enable the estimated motion vectors to compensate for the scene changes between frames caused by objects' motion. However, this approach may not be suitable to provide practical and cost saving solutions, considering the complexity of such algorithms.

4.5 Experimental Results

The proposed real-time 3D surface reconstruction approach is evaluated in this section. The same image sequences used for the segmentation evaluation were reused for the experiments. The ambient illumination of the sequences were successfully eliminated after applying the DoubleFlash method discussed in Section 2.3, while the image distortion caused by the optical lens was corrected using the camera parameters provided by the camera calibration. The fixed pattern noise as well as the background components of the input sequences were also suppressed by subtracting the reference background from the sequences. *Three* sequentially captured image frames were finally used as an *input triplet* for the photometric stereo method for surface reconstruction. These input triplets were associated with the segmentation results of the shadow-suppressed images for specifying the region of interest to reconstruct. The *overlapped* region of the three segmentation results was taken into account for the surface reconstruction. Assuming a Lambertian surface, the surface normals were computed from each pixel while some extremely erroneous vectors detected by simple thresholding were replaced by the average of the neighbour surface normals. By minimising the Wei energy function presented in Equation 4.10, the object surface was integrated based on the estimated surface normals in the region of interest.

4.5.1 Reconstruction examples

Some typical surface recovery examples of the different occupant classes as well as the needle map representations of their surface normals are shown through Figure 4.6, 4.7, and 4.8. Figure 4.6(f) shows the segmentation result applied on the ShadowFlash image (e) composed of the *input quadlet* (a)-(d). The object boundary information was used for determining valid surface normals for the surface integration process by discarding the normal vectors out of the region of interest. As a consequence of flaws in the segmentation results, the surface normals on the fringe of the object tended to deviate as shown in Figure 4.6(g). After scaling the relative depth information by a constant, the successfully reconstructed surface of the FFCS class is presented in Figure 4.6(h). Although the recovered surface experienced some distortions caused by the non-linear characteristics and the perspective projection of the imager, the recovered object surface provided the sufficient information of the shape of the given object. In Figure 4.6, 4.7 and 4.8(i), the surface reconstruction results of three occupant classes are shown after rotated by 90 degrees to display the surfaces on the x-z image plane. The surfaces computed by the proposed system did not provide superior details of the object shapes to the stereo vision based system. However, the apparent difference of the recovered surfaces between the occupant classes nonetheless provided sufficient information for distinguishing those classes.

4.5.2 Evaluation of the reconstructed surface accuracy

The examples for evaluating the surface reconstruction sensitivity to different depths is demonstrated in Figure 4.9. Indeed, the quantitative evaluation of the reconstructed surface accuracy computed by any shading-based shape recovery methods is not simple as the methods only provide *relative* depth. Two input triplets having different depth properties were acquired, and the surfaces reconstructed from these inputs were used for a depth comparison after being scaled by the same constant factor.

Figure 4.9(a) shows the ShadowFlash results of the input triplets in the situations being a child seat occupied by a *baby* and a child seat with a *baby holding a toy*. The needle map representation of the surface normal vectors and the surface reconstruction results based on the estimated surface normals are presented in Figure 4.9(b) and (c), and the 90-degree rotated versions of the results are compared in Figure 4.9(d). Similar to the former surface reconstruction examples, the depth difference caused by the use of different objects was successfully reflected in the reconstruction results despite the limited descriptions of the surfaces.

Another example is shown in Figure 4.10 to evaluate the accuracy of a reconstructed object surface. Two different situations are setup depending on a *sphere-shaped* object and a child seat under the identical illumination conditions. The input triplets are displayed on the left side of Figure 4.10, where each of the triplet elements is superimposed on the same image plane using different colour space. Although the surface of the sphere object is significantly distorted as a consequence of the inaccurate surface normal estimation, the depth difference between the two examples are apparent while the overall shapes of the child seats are successfully recovered as shown in the right side of Figure 4.10.

4.5.3 Surface reconstruction of temporal sequences

An example sequence of a vehicle passenger leaning forward is shown in Figure 4.11(a). To reflect the general tendency of vehicle passengers' movements, the behaviour of the passenger was not under control during the sequence acquisition. Although the sequence was captured at 30 Hz, only every fifth frame of the sequence is displayed in this example. As shown in Figure 4.11(b), the surface normals on the passenger's head often contained severe errors due to the estimation ambiguity caused by the inhomogeneous surface reflectivity. The intensity values of the input triplet at the outline of the object were extremely sensitive to the deformation of the object, and such sensitivity often caused the false integration of the surface. In Figure 4.11(c) and (d), the surface integration results of the given sequence from the different views are presented. Despite the considerable time delay between input frames, the recovered surfaces maintained the rough shapes

of the passenger's head and shoulder. The results are encouraging as they suggest the possibility of extension of the proposed system to the passenger out-of-position detection system.

4.6 Precis

By exploiting the pre-existing active illumination hardware, the photometric stereo method is employed as the surface reconstructing technique of the proposed system. The suitability of the proposed approach to the low-cost real-time 3D imaging applications is compared to that of the classical stereo vision techniques under the assumption of operation environments characterised by objects with no significant textures, close proximity between the imager and the target objects, and geometrical constraints imposed by limited space. The surface normal estimation with three differently illuminated input images is discussed followed by the introduction of the Wei algorithm [104] based on the global minimisation of several constraints. Comparison of the proposed system with stereo vision techniques is discussed in terms of ease of implementation, robustness to noise, and economic aspects.

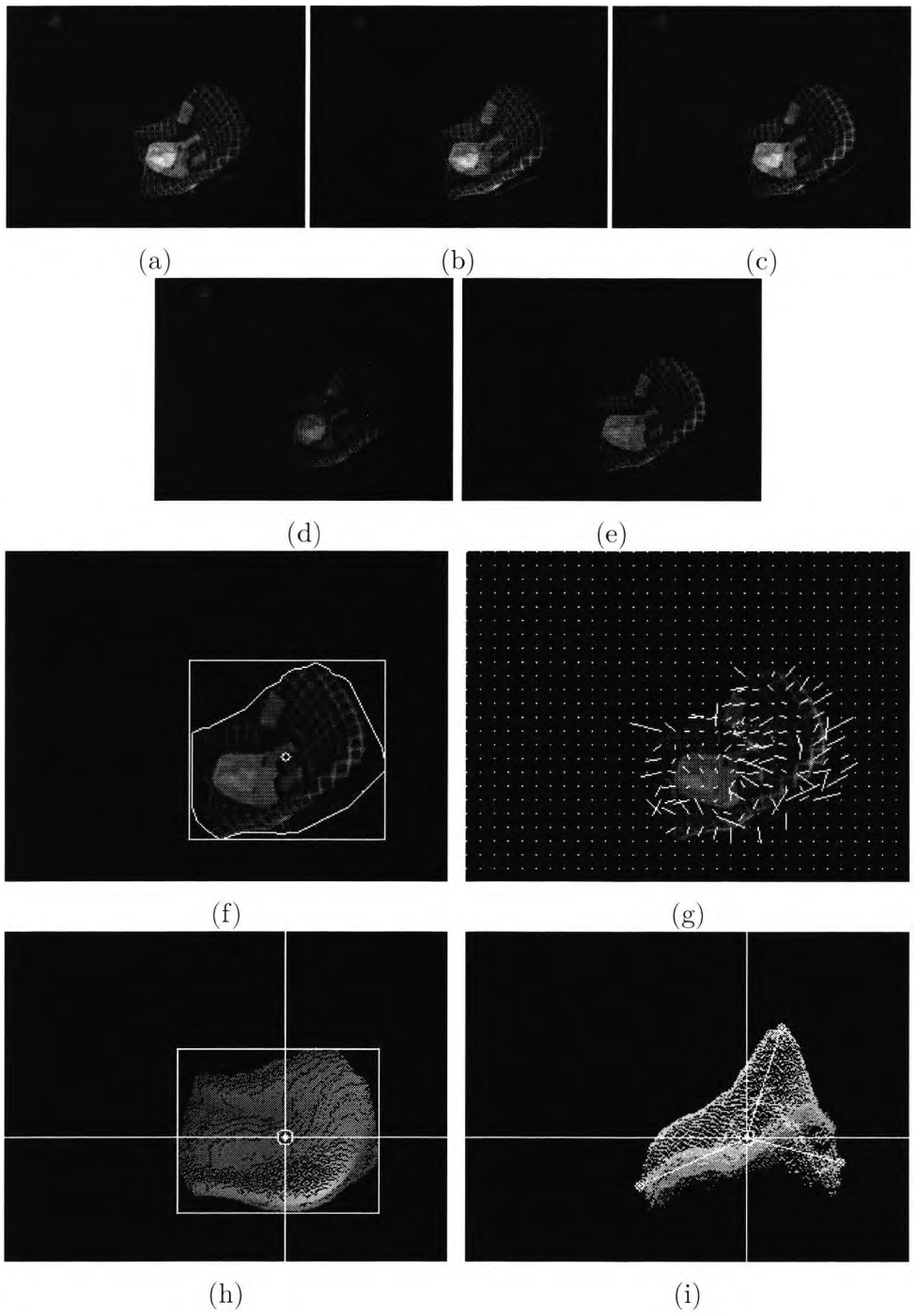


Figure 4.6: Examples of the surface reconstruction of a FFCS class: (a)-(d) the input quadlet, (e) ShadowFlash result, (f) segmentation result, (g) needle map, (h) and (i) surface reconstruction results viewed from the top and rotated 90 degree about y-axis.

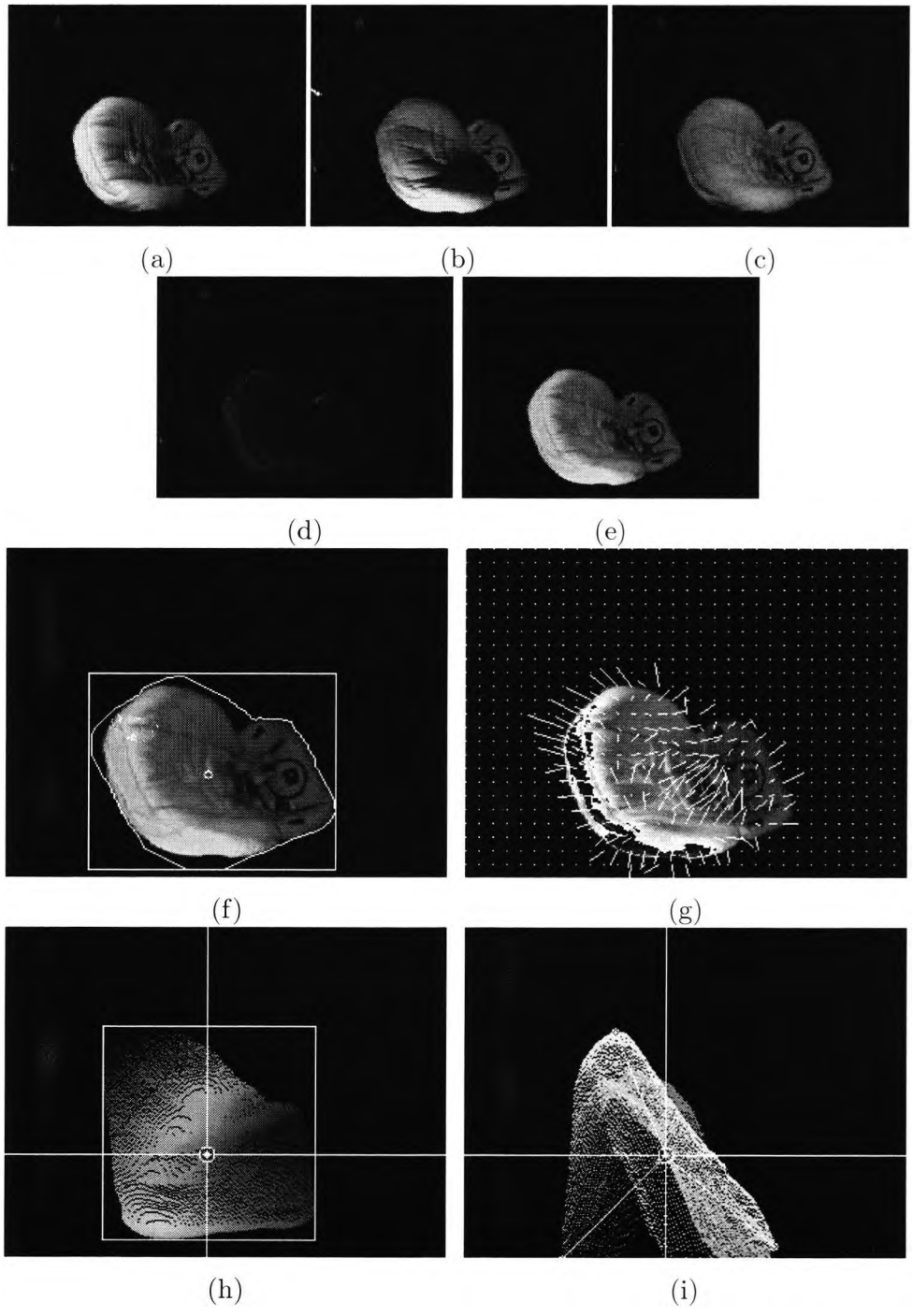


Figure 4.7: Examples of the surface reconstruction of a RFCS class: (a)-(d) the input quadlet, (e) ShadowFlash result, (f) segmentation result, (g) needle map, (h) and (i) surface reconstruction results viewed from the top and rotated 90 degree about y-axis.

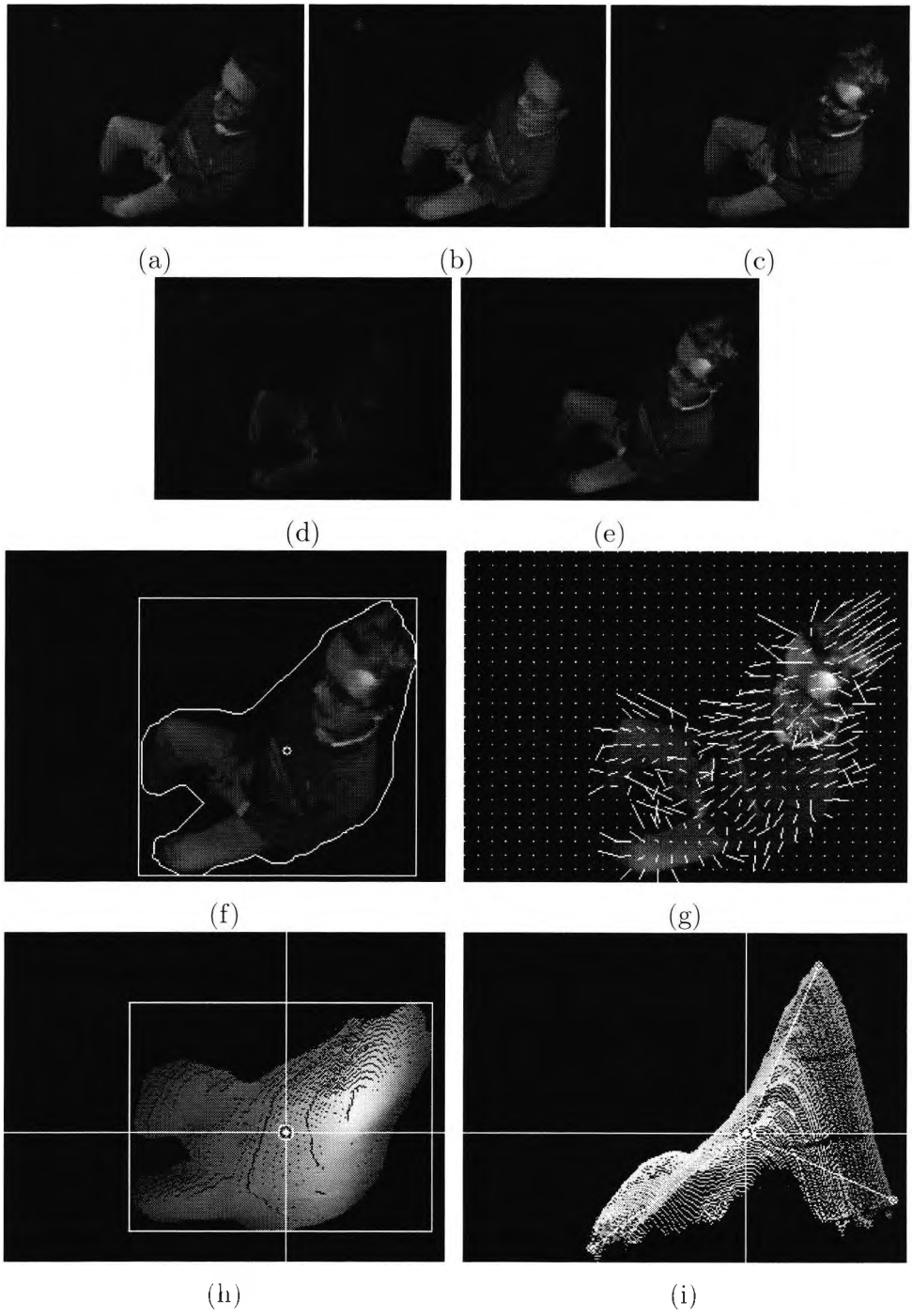


Figure 4.8: Examples of the surface reconstruction of an adult class: (a)-(d) the input quadlet, (e) ShadowFlash result, (f) segmentation result, (g) needle map, (h) and (i) surface reconstruction results viewed from the top and rotated 90 degree about y-axis.

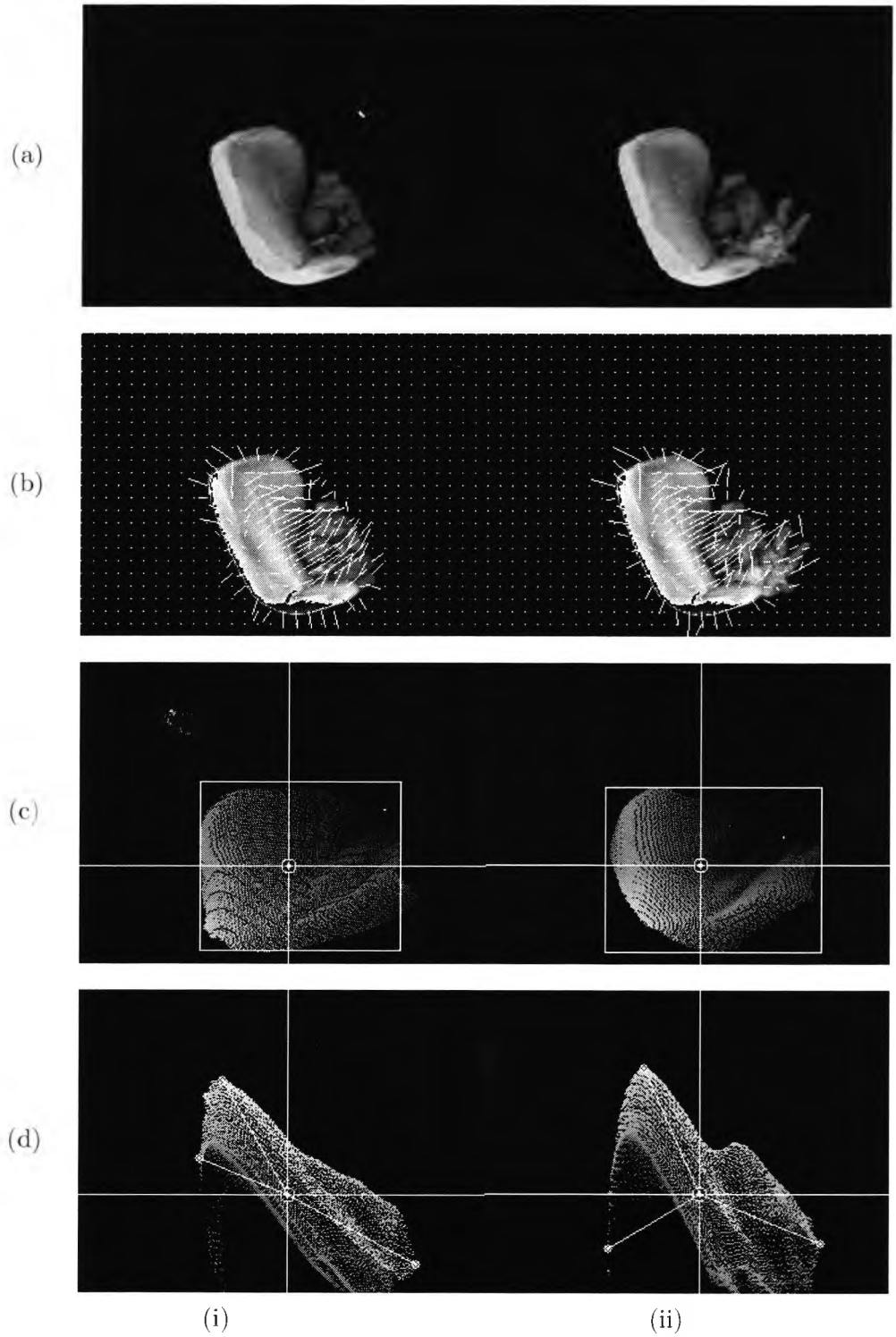


Figure 4.9: Evaluation of surface reconstruction sensitivity to different depth: (a) the original frames, (b) needle map representations, (c) and (d) surface reconstruction results viewed from the top and 90 degree rotated, respectively. (i) a child seat occupied by a baby and (ii) a child seat with a baby holding a toy.

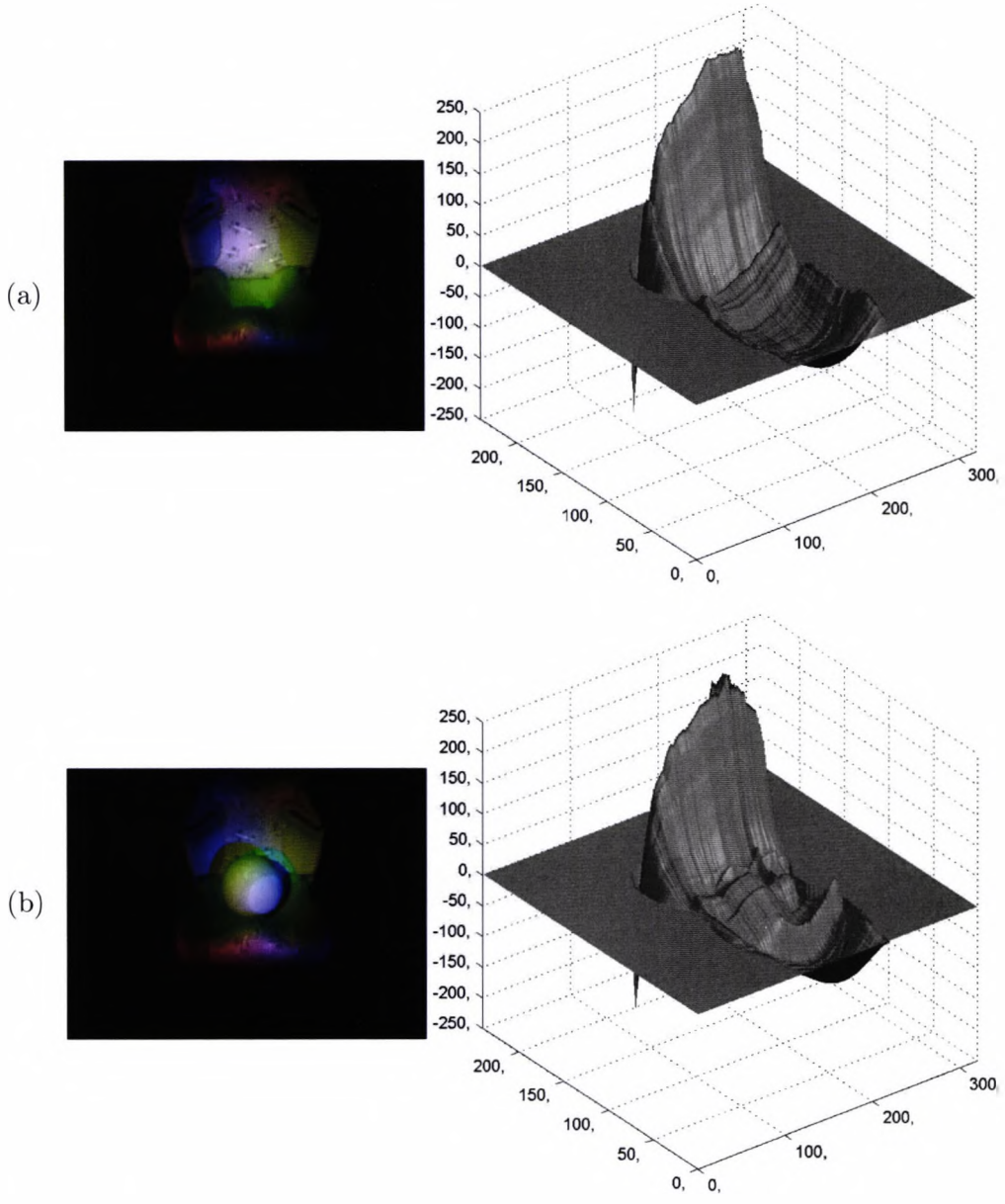
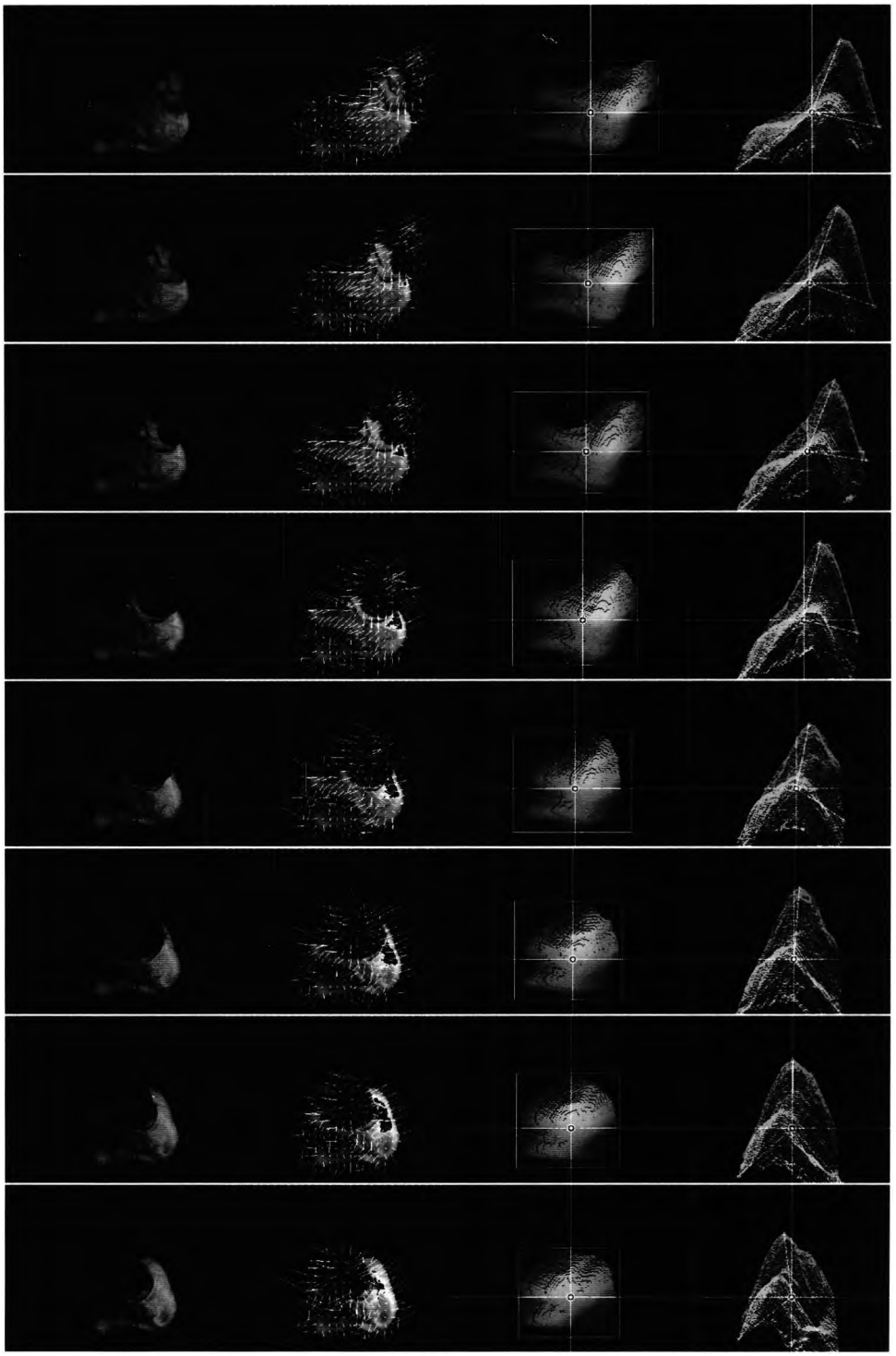


Figure 4.10: Comparison of the reconstructed surfaces with different depths: the accuracy of the reconstructed surface can be compared using (a) an empty seat and (b) the same seat occupied by a sphere-shaped object.



(a) (b) (c) (d)

Figure 4.11: An example sequence of the 3D surface reconstructed from an adult class: (a) the original input sequence, (b) the needle maps, (c) and (d) the sequence of the reconstructed surface viewed from the top and rotated 90 degree, respectively.

Chapter 5

Classification

5.1 Introduction 102
5.1.1 Motivation 102
5.1.2 State of the art 102
5.1.3 Summary 104
5.2 Feature selection 104
5.2.1 Extended Gaussian image 104
5.2.2 Surface depth 105
5.2.3 Spread axes information 105
5.2.4 Relative position of the upper extremum 107
5.2.5 Volumetric ratio and compactness 107
5.2.6 Other 2D geometric information 108
5.3 Classifier design 108
5.3.1 Occupant class assignment and operation assumption 108
5.3.2 System design requirements 110
5.3.3 Neural networks 110
5.3.4 Implementation 114
5.4 Precis 116

5.1 Introduction

5.1.1 Motivation

One of the tasks most machine vision systems must accomplish is classification. Pattern classification could be summarised as the categorisation of some input data into identifiable classes via the extraction of significant features or attributes of the data from a background of irrelevant details. Pattern classification has found various applications including character recognition, fingerprint identification, minefield detection, vehicle occupant classification, etc.

The basic idea underlying pattern classification is to extract *features*, which are measurements of quantities considered useful in distinguishing members of different classes. Measurements of different features are then adjoined to form a *feature vector*, and the *classifier* assigns an object to a category by utilising the abstraction provided by the feature vector representation about the object of interest. Finally, the information obtained from the image of an object can be used to identify a point in some multidimensional feature space [39].

The degree of difficulty of the classification problem depends mainly on the *variability* in the feature values for objects in the same category relative to the difference between feature values for objects in different categories. The variability of feature values for objects in the same category may be due to complexity or noise. Noise interferes with all non-trivial decision and pattern recognition problems in some form. Noise can be defined as any property of the sensed pattern which is not due to the true underlying model but instead to randomness in the real world [21]. For example, a large number of highly complex transformations arise in pattern recognition. Transformations such as non-rigid deformations which arise in three-dimensional object recognition are far more severe. A good example is the radical variation in the image of a constantly moving vehicle occupant. Similarly, variations in illumination or the complex effects of cast shadows may need to be taken into account.

5.1.2 State of the art

Like other machine vision applications, the efficient representation of a feature vector as well as the selection of a appropriate classifier are of great significance for an occupant detection system. In [55], Krumm proposed a *three*-class occupant classification method based on *template matching*. Principle components analysis was employed to provide a set of *prototype* images from the preprocessed image data. A number of *invariant coefficients* were generated by the eigenvectors extracted from both the intensity and disparity data and used for image comparison based on the *nearest neighbour*

classifier. Relatively high classification performance was achieved based on a limited idealised image set. However, the simple features sensitive to the variation of environmental conditions lowered the feasibility of such a system.

Legendre moments representation of the edge image obtained from a single grey-scale camera was employed as the feature set by Farmer in [23]. In order to solve a *four*-class classification problem, *multiple* k-*nearest neighbour-based classifiers* were combined by taking the average of the probabilities from each of those classifiers. The final decision was made by choosing the class with the lowest average distance to its k-nearest neighbours based on the Manhattan distance metric. Although the classification results were encouraging considering the large intra-class variation, the ambiguity of the feature set caused by the two-dimensional projection of three-dimensional deformable objects might result in insufficient system reliability in reality. Furthermore, shape based image matching techniques have been used only in limited capacities and are still not mature enough to provide a high degree of automated classification process.

The *eight*-class occupant detection system proposed by Marín-Hernández in [62] utilised the global occupancy description built from the percentage of 3D points in the 3D surface image for classification. A recursive algorithm is proposed to find the minimum number of 3D points which define the external surface of the passenger. The areas between the passenger and airbag were divided into several regions, and the number of the 3D points counted in each region was used for composing a feature vector. The classification was made by the method of the k-*nearest neighbours* based on a relatively small database. The contribution of this work was the introduction of the *volumetric density* in three-dimensional representation deduced from the geometric information of a given occupant. However, the 3D points density became greater in the surfaces closer to the imager because of the severe perspective caused by the wide-angle lens configuration. Furthermore, the system performance depended entirely on the reliability of the provided surface information which could be easily distorted by various noise sources such as mechanical vibration, illumination change, partial occlusion, etc.

Koch introduced an occupant detection system using a single monochrome camera with active illumination in [51]. Based on the *three*-dimensional shape representation of occupants, several useful geometrical features were defined such as *spread angles* and *blob proportions*. *Six* occupant classes were assigned to distinguish the objects in the moderate size of image data collected in the real vehicle environment, and the theory of *fuzzy logic* is employed for a classifier. The experimental results successfully showed the feasibility of an occupant classification system based on 2D image processing with high performance. The major problem of this work was the lack of consideration for the perspective distortion caused by the projection of three-

dimensional objects into the two-dimensional image plane. By ignoring the influence of the projection distortion, the classification rate was significantly reduced as a result of the increased ambiguity between the occupant classes.

5.1.3 Summary

In this chapter, the problem of categorising vehicle passengers into several predetermined occupant classes is addressed. The aim of this chapter is to (1) minimise the system implementation cost by decreasing the role of the classifier with computationally efficient and economic features, and (2) obtain high reliability as well as performance independence from the severe automotive environmental conditions. A number of useful features based on both *two-* and *three*-dimensional geometric attributes of possible vehicle occupants are suggested, in which each feature is designed to discriminate at least one occupant class from the other two classes. The dimension of the feature vector is limited to 29 dimensions by taking into account beforehand the *curse of dimensionality* problem. To reflect the dynamic properties of the occupants, a neural network with the *partially recurrent structure* is proposed as the classifier designed to solve a *three*-class problem. Two *tapped delay lines* are employed for stabilising the erroneous fluctuation of feature values and for providing time-varying weights resulting from the confidence change of the extracted features caused by the occupants' motion.

5.2 Feature selection

The choice of *distinguishing features* is a important design step and depends on the characteristics of the problem domain. Although seeking the distinguishing features invariant to any irrelevant transformations of input is an essential task to make the job of the classifier trivial, it was still difficult to find apparent features which clearly discriminate all the classes. Therefore, each feature is designed to specify at least one class from the others. For example, the occupant size could be used for distinguishing an adult from the child seat classes. The definitions of the proposed features are discussed in this section. Another important design factor to take into account is to guarantee the invariability with respect to possible translations which could be produced by the movable passenger seats.

5.2.1 Extended Gaussian image: 4 dimensions

The *extended Gaussian image* (EGI) is a histogram of the surface normals computed over a discretised Gaussian sphere. The surface normals are easily obtained during surface reconstruction. As it can be expected that the rear-facing child seat should have a different aspect of its surface direction from

the ones from the other two classes, the surface normals can be used as a key feature. The object surface is divided into surface patches based on the similarity of the surface normal primitives using *quadtree subregioning*. At each surface patch, an averaged surface normal vector is calculated from the surface normal primitives belonging to the patch. Consequently, the influence of the erroneous normal vectors caused by various noise sources such as the *frame delay* and trivial *reflectance model* can be minimised. Finally, the EGI histogram is divided into bins of 90 degrees each, and the number of the averaged normal vectors belonging to each bin are calculated. Normalised by the sum of the total average surface normals, the *four*-dimensional EGI features invariant to translation are defined.

5.2.2 Surface depth: 4 dimensions

The profile of the depth information projected from the top of an object is also used as a feature. The *relative* depth computed as a result of the photometric stereo method cannot be directly used for the quantitative analysis of the object shape. Therefore, the estimated depth values of each input frame are normalised to the numbers between *zero* and *one*. Since the camera coordinate system differs from the world coordinate system, a rotational transformation is performed with a given rotation matrix $\mathbf{R}$ representing three partial transformations (*pan*, *tilt* and *roll angles*) in order to provide a depth profile projected along the z-axis of the world coordinate system. A brief illustration of changing the view point is shown in Fig.5.1. After the rotational transformation, the object surface is divided into *four* equally-spaced regions while from the average of the normalised surface depth is computed from each region.

5.2.3 Spread axes information: 9 dimensions

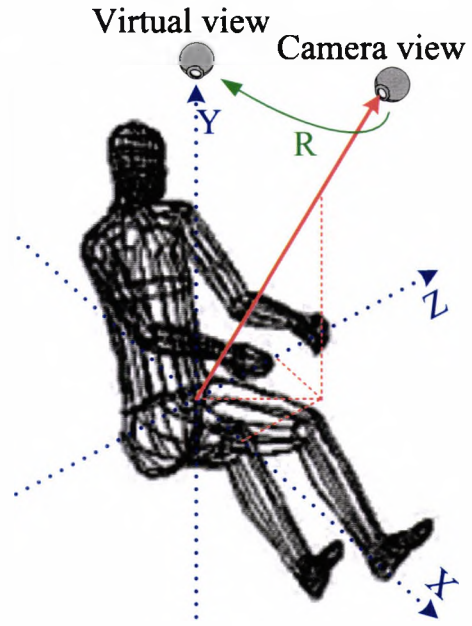
With the successful recovery of the object surface, three extrema E_1, E_2 and E_3 on the surface are defined in three dimensional space as shown in Figure 5.2. These extrema are used for defining the following useful features:

Spread axes The *spread axes* are the lines between the centre of gravity of the object volume and the extrema.

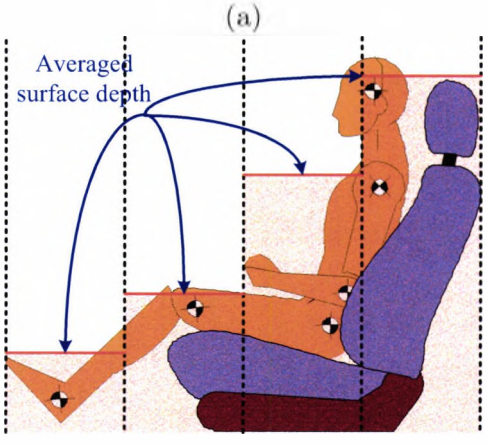
Spread angles The *spread angles* α_{Z1}, α_{X2} and α_{X3} are defined as the angles between the spread axes and the coordinate system

Relative spread angles The *relative spread angles* β_{12}, β_{13} and β_{23} are the angles between the spread axes themselves.

These two angle characteristics and the lengths of the spread axes are used as *nine* key features for the classifier. Since these three feature sets are correlated with the centre of gravity, they become invariant to any translation transformation. The features become independent of the object size as well



(a)



(b)

Figure 5.1: Features from depth information: (a) the camera calibration provides the rotational matrix **R** with respect to the world coordinate origin. In principle, all the three-dimensional features are rotationally transformed in order to make them correctly viewed from the top, and (b) the four equally spaced regions are defined on the object surface and the average depth levels of these regions are utilised as 4-dimensional features.

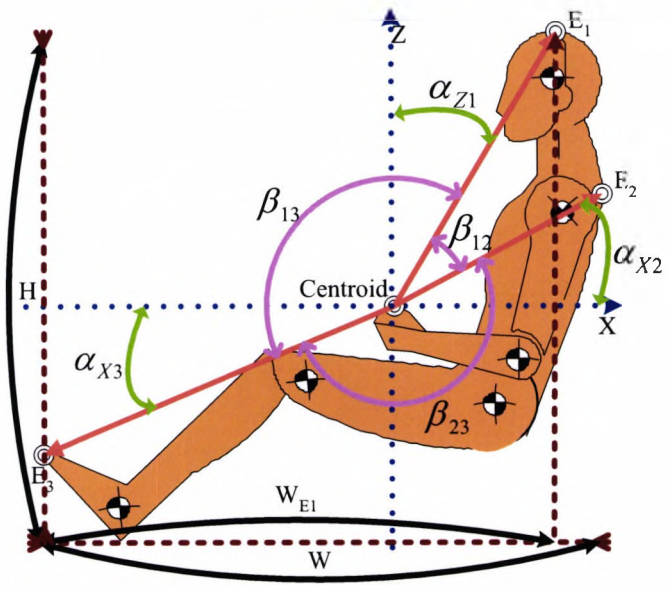


Figure 5.2: Spread axes information: the extrema E_1, E_2 and E_3 are defined as a most upper, most front (left) and most rear (right) point on the recovered surface, respectively.

as its absolute position by being normalised by the maximum value of each feature set. A few examples are shown in Figure 4.6, 4.7, and 4.8(i).

5.2.4 Relative position of the upper extremum: 1 dimension

The relative position of the upper extremum E_1 along the x-axis could be a good clue to specify the rear-facing child seat class against the other two classes. As shown in Figure 5.2, the relative position P_{E_1} is simply defined as

$$P_{E_1} = \frac{W_{E_1}}{W} \tag{5.1}$$

where W and W_{E_1} are the width of the object and the distance along the x-axis between the $E1$ and $E3$, respectively. This feature is independent of translation since the feature value is only correlated with/normalised by the width of the target object.

5.2.5 Volumetric ratio and compactness: 2 dimensions

As it is not possible to recognise what happens behind the object, it is difficult to define the absolute *volume* of the object. Even if the assumption is made that the object has a flat back side, the volume of the target may still be extremely sensitive to the segmentation result. For example, a few pixels of errors in extracting the object boundary may result in a significant overestimation of the volume due to a couple of improper *z-layers*

produced by the erroneous boundary during the surface integration process. Consequently, the *ratio* of the three-dimensional surface area to the two-dimensional boundary area is defined as the *volumetric ratio*, which should increase as the volume of the object expands. Assuming a flat back side, the proportion, or *compactness*, of the object volume to a hexahedron enclosing the object could also provide robust estimation of its volume.

5.2.6 Other 2D geometric information: 9 dimensions

In [40], Hu derives a set of seven functions which make use of the central moments of an image blob. Their output is independent of any translation, rotation or mirror image of a particular blob, and they can be used in conjunction with both the image blob itself and the edge-processed contour image. Hu's equations are based on the *uniqueness theory of moments*: the infinite sequence of moments m_{pq} is uniquely determined by the joint function $f(x, y)$; conversely, the function $f(x, y)$ is uniquely determined by an infinite sequence of moments m_{pq}. For a digital image of size (N, M) the $(p + q)$-th order moments m_{pq} are calculated as

$$m_{pq} = \frac{1}{NM} \sum_{x=1}^{N} \sum_{y=1}^{M} f(x, y) x^p y^q \tag{5.2}$$

for $p, q = [0, 1, 2 \cdots)$. Similarly, the normalised central moments of a digital blob image are inherently translation independent. The definition of the normalised central moments for $p, q = [0, 1, 2 \cdots)$ is

$$\mu_{pq} = \frac{1}{NM} \sum_{x=1}^{N} \sum_{y=1}^{M} f(x, y)(x - \bar{x})^p (y - \bar{y})^q \tag{5.3}$$

where $(\bar{x}, \bar{y})$ denotes the centroid of the contour.

In general, gross image shape is well represented by the lower-order moments, and high-order moments reflect only the subtleties of a silhouette or boundary image. Nearly all work with moment invariants, including the normalised and Hu's central moments, depend only on moments of order zero to three. Since minor differences in outlines of the vehicle passengers are not significant, only *three* low-order components of both *normalised central moments* and *Hu moments* are selected as features, along with the *width*, *height* and two-dimensional *area* of the object boundary.

5.3 Classifier design

5.3.1 Occupant class assignment and operation assumption

The goal of most *smart* airbag systems is to categorise the objects in a passenger seat into *six*-predetermined classes for adaptive airbag deployment

as discussed in Section 1.2.2. In reality, it is difficult to design an ultimate system which perfectly discriminates *all* types of potential occupants that could be presented within a vehicle. Such a system should be completely independent of the geometrical variations of objects, robust to severe illumination conditions, and invariant to transformations including objects' deformation. This may be realised only if either an infinite amount of data is provided or several of the available safety technologies work together.

The less ideal but more *attainable* solution could be provided by reducing the problem domain at hand. The complexity of a classifier mostly depends on the *number of classes*, which is determined based on both the analytic purposes of the target system and the quality of available data. The increase of target classes requires the larger dimensions of plausible features along with the more sophisticated pattern recognition theory. And the demand for a large number of samples grows exponentially with the dimensionality of the feature space. This limitation is called the *curse of dimensionality* introduced in [4], and severely restricts the practical classification applications. The fundamental reason for the curse of dimensionality is that high-dimensional functions have the potential to be much more complicated than low-dimensional ones, and that those complications are harder to discern [21]. Moreover, if little or nothing in the way of data reduction is provided, this leads to severe requirements for computation time and storage.

Consequently, the number of occupant types to be classified in this work is reduced based on the following assumptions: (1) there is no need for detecting an empty seat (NPOS) or other unknown objects which do not fit to the other occupant classes (ODFC) for a safety reason. (2) Since a *passenger out-of-position detection system* takes charge of discriminating between the *person out-of-position* (POOP) and *person in correct seating position* (PCSP) classes as shown in Figure 1.4, only the superclass of those two person classes called *adult* is provided by the proposed occupant detection system. Finally, the number of occupant classes is limited to *three* by distinguishing only two child seat classes and an adult class: *adult*, *forward-facing child seat* (FFCS) and *rear-facing child seat* (RFCS).

Since a change of the occupant type is unlikely during driving, it is sufficient to perform the classification only at the beginning of operation in most cases, unless any dramatic change in the field of view occurs (see Figure 1.4(b)). Alternatively, the classification can be performed periodically if there are still sufficient processing power and hardware resources remaining for the passenger out-of-position detection system. The average processing time required to make a decision for existing vision-based occupant detection systems is around three seconds. A goal of the proposed system is to reach a decision within this aforementioned *three seconds* time frame. During this

time, this system should be capable of processing of 90 frames with a frame rate of 30 Hz prior to making a final decision.

5.3.2 System design requirements

Several requirements has to be satisfied to realise a practical occupant detection system. The design of the proposed classifier should be flexible and adaptable to take into account future child seat designs. Some practical issues which may arise include (1) the necessity of frequent system updates with new data sets and (2) the flexibility to add further types of classes to the classifier structure to accommodate potential occupant class variation. However, it is unlikely to recall the products whenever an update is necessary. Furthermore, no modification of the system structure will be allowed after the completion of the system design due to the vehicle safety regulations. Accordingly, system updates should be restricted to changes of the internal parameters by minimal remote data transmissions, and minor modification to non-critical system components by means of officially approved firmware updates.

5.3.3 Neural networks

A generic *neural network* (NN) can be described as a computational system consisting of a set of highly interconnected processing elements called *artificial neurons*, which process information as a response to external stimuli [19, 21]. An artificial neuron is a simplistic representation that emulates the signal integration and threshold firing behaviour of biological neurons by means of mathematical equations. The network adapts to the given data by changing the *connection weights* by an amount proportional to the difference between the desired output and the actual output. Neural networks have seen an explosion of interest over the last few years, and are being successfully applied across an extraordinary range of problem domains, in areas as diverse as finance, medicine, engineering, geology and physics. Indeed, wherever there are problems of prediction, classification or control, neural networks are being introduced.

A neural network can be one of the available solutions for the problem addressed in Section 5.3.2. The *adaptability* to the future child seat variation and the potential for the increasing occupant classes could be achieved by changing the interconnection weights between neurons. Neural networks are also more *robust* at data analysis than statistical methods because of their ability to handle small variations of parameters and noise. Various high performance *neural network processors* available in the market should make the implementation of real-time embedded systems simple. Despite the relatively slow learning process, a *fast* classification is another merit of employing neural networks.

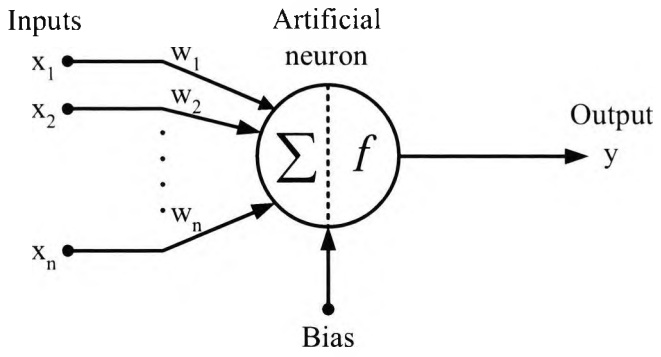


Figure 5.3: The structure of a neuron: inputs from one or more previous neurons are individually weighted, then summed. The result is scaled between 0 and 1 by an activation function, and the output value is passed on to the neurons in the next layer.

Introduction

Figure 5.3 shows an illustration for a basic processing unit in a neural network. A neuron receives a number of inputs x either from original data, or from the output of other neurons in the neural network. These input signals are passed between neurons over connection links where each connection link has an associated weight w multiplying the transmitted signal. Each neuron has an internal state called its *activation* which eventually becomes the output of the neuron. An *activation function* f is applied to the sum of weighted input signals to determine its activation signal while the *bias* b allows the activation to change independently of the inputs. Many different functions may be used as activation functions depending on the desired output characteristics. Typically, a neuron sends its activation as a signal to several other neurons. A neuron can send only *one* signal at a time. However, a signal may be broadcast to *several* other neurons. The mathematical expression of a typical artificial neuron is as follows:

$$y = f\left(\sum_{i=1}^{N} w_i x_i + b\right) \tag{5.4}$$

A key feature of neural networks is an iterative learning process in which a set of sample data is presented to the network one at a time, and the weights associated with the input values are adjusted each time. After all cases are presented, the process often starts over again. During this learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label of input samples. Advantages of neural networks include their high tolerance to noisy data, as well as their *generalisation* ability to classify patterns on which they have not been trained. *Generalisation* refers to the ability of a neural network, having learned the essential

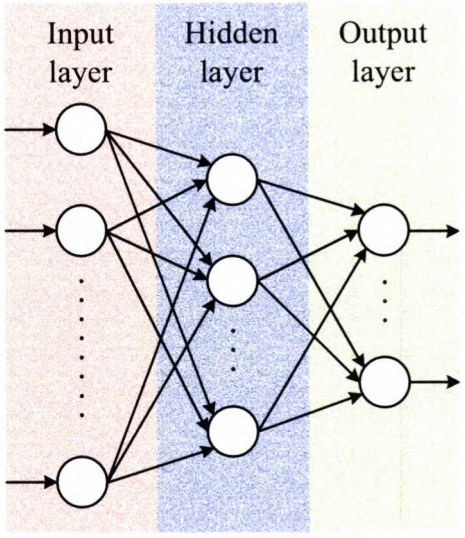


Figure 5.4: A typical example of a multi-layer neural network: the typical back-propagation network has an input layer, an output layer, and at least one hidden layer.

information content of training data, to achieve reasonable performance for test data not seen before which is drawn from the same input space.

Network topologies

There are two principal neural network topologies which define how data flows between the input, hidden, and output processing units: *feed-forward* and *recurrent networks*.

Feed-forward networks (FNNs) A neural network where the data flow from input to output units is strictly feed-forward, is called as a *feed-forward* network. The earliest kind of neural network is a two-layer perceptron network originally introduced by Papert in [65]. This perceptron network consists of a single layer of output nodes while the inputs are fed directly to the outputs via a series of weights. The crucial problem of the perceptron is that this network is only capable of learning linearly separable patterns by allowing only one layer of adaptive weights. The exclusive-or (XOR) function is a classical example of a pattern classification which is *non-linearly separable*. Feed-forward networks with more than two layers, also called as *multi-layered perceptrons*, overcome this limitation because they are able to adapt *multi-layers* of weights by using more sophisticated learning rules. *Back-propagation*, popularised by Rumelhart in [85], is one of the most popular and effective learning models for multi-layered networks. The power of back-propagation lies in its ability to train hidden layers and thereby escape the restricted capabilities of the perceptrons by providing a complex

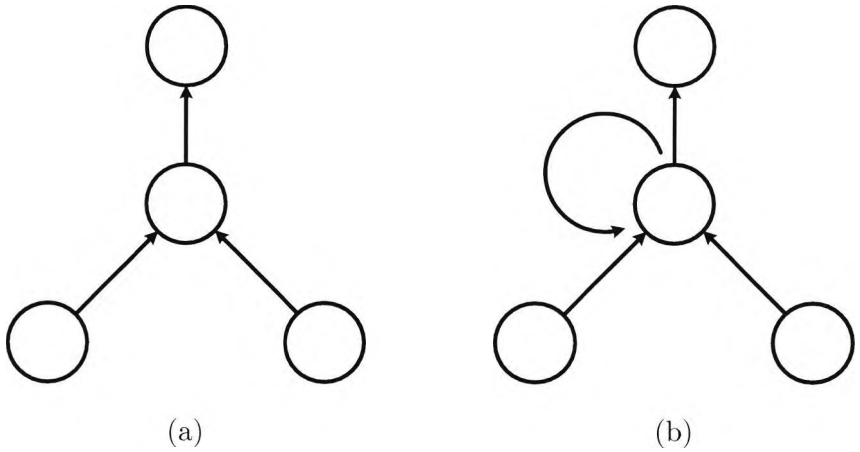


Figure 5.5: The conceptual illustrations for two network topologies: (a) a *feed-forward* network and (b) a *recurrent* network with an additional connection from the hidden unit to itself.

non-linear decision boundary. Figure 5.4 illustrates a typical structure of three-layered back-propagation networks.

Recurrent networks (RNNs) In cases where neural networks deal with data in the temporal domain, the most common architecture is a recurrent neural network with internal feedback connections which makes the system biologically more plausible. Basically, a recurrent neural network is a modification to the feed-forward architecture for temporal data processing. The conceptual difference between the feed-forward and recurrent networks is illustrated in Figure 5.5. In recurrent networks, information about past inputs is fed back into and mixed with the current inputs through recurrent or feedback connections for hidden or output units. In this way, the neural network contains a memory of the past inputs via the activations. These recurrent networks can have an infinite memory depth and thus find relationships through time as well as through the instantaneous input space. Most real-world data contains information in its time structure. The recurrent networks have an dynamic internal state which is essential for many temporal processing tasks. Therefore, they are computationally more powerful than other adaptive models such as Hidden Markov Models, Support Vector Machines, and feed-forward networks [13, 71, 35]. In principle, the recurrent networks can implement almost arbitrary sequential behavior, which makes them promising for various applications such as adaptive robotics, speech recognition, music composition, attentive vision, etc.

Fully recurrent networks acquire their dynamic properties by providing two-way connections between all processors in the neural network. Unlike feed-forward network variants which have a *deterministic* time to produce

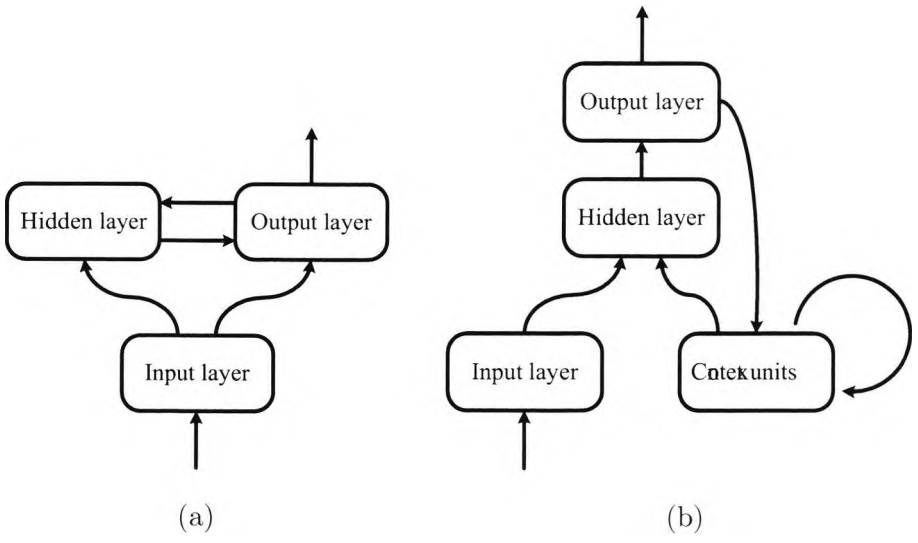


Figure 5.6: Recurrent network topologies: examples of (a) a fully recurrent neural network and (b) a partially recurrent network.

an output value based on the time for the data to flow through the network, the fully recurrent networks can take an *indeterminate* amount of time.

Partially recurrent neural networks are feed-forward networks that include feedback connections to a set of units called *context units*. A *context unit* is basically an internal state memory which remembers past activity of the network. This structure of recurrence compromises the system complexity between a feed-forward network and a fully recurrent network due to the capability of using the popular *back propagation* training algorithm. In [44], Jordan proposed three-layer back-propagation networks, with the addition of feedback connections from the output layer to its context units. This internal feedback loops make the *Jordan networks* capable of learning to recognise and generate temporal patterns, as well as spatial patterns. This makes the Jordan networks useful in such areas as signal processing and prediction where time plays a dominant role.

5.3.4 Implementation

Considering that the proposed features did not reflect any dynamic properties of the passenger, it was necessary to construct a classifier model which is able to handle and classify temporal series. Therefore, trained in a supervised way, a *partially recurrent network* proposed by Jordan [44] is employed with the support of two *tapped delay lines*.

The network is designed to have 29 input units and 3 output units according to the dimension of the extracted feature vector as well as the number

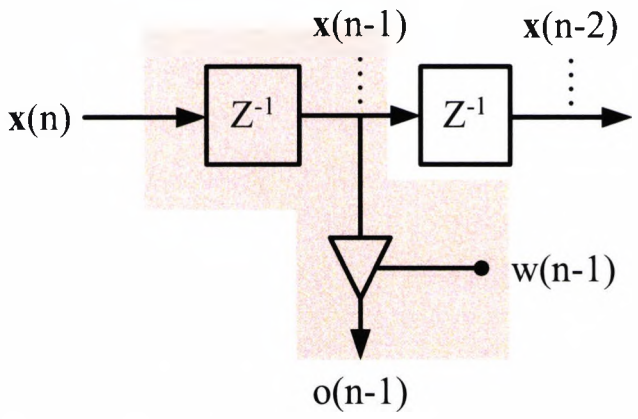


Figure 5.7: A delay line with one tap: the output of the delay line o is obtained from the delayed signal $\mathbf{x}(n-1)$ multiplied by the weight $w(n-1)$.

of occupant classes. Since the system complexity is dependent on the number of units, the optimal structure of a neural network is expected to have the minimal number of hidden units sufficient to achieve the desired error value on the training set. In most situations, there is no way to determine the best number of hidden units without training several networks and estimating the generalisation error of each. Insufficient number of hidden units can cause high training error as well as high generalisation error due to the under-fitting problem. Conversely, too many hidden units may yield low training error but the system may still have high generalisation error due to over-fitting and high variance [86, 6]. For the proposed system, the optimal number of the hidden units is experimentally obtained to be 15 by training the network with a different number of the hidden units.

A *tapped delay line* is a delay memory providing access to its contents at arbitrary intermediate delay length values. Each tapped delay line improves the accuracy of overall classification performance by filtering the noise components in the stream of either feature vector (input) or classification result (output). As discussed in Section 4.4.5, the occupants' motion can produce erroneous patterns by distorting the reconstructed object shapes despite a fast frame rate. A *time-varying weight* $w(n)$ based on the amount of *motion* is used for minimising the undesirable influence of the surface distortion caused by motion. For the simplicity, the amount of motion between the adjacent frames is measured by counting the number of pixels which experience brightness changes beyond a certain threshold. The number of the motion pixels are then normalised by the sum of the motion pixels occurred in the delay lines. A delay line with a single tap is shown in Figure 5.7. Assuming that the length of a delay line is N, the output of the delay line $\mathbf{o}$ is computed as

$$\mathbf{o} = \mathbf{X} \cdot \mathbf{w}^T$$

where $\mathbf{X}$ and $\mathbf{w}$ are the vectors of the input pattern matrices $\mathbf{x}$ and time-varying weights w, composed of their time series for the interval $[0, N]$:

$$\begin{aligned} \mathbf{X} &= (\mathbf{x}(n), \mathbf{x}(n-1), \cdots, \mathbf{x}(n-N)) \quad \text{and} \\ \mathbf{w} &= (w(n), w(n-1), \cdots, w(n-N)) \,. \end{aligned}$$

The average of the past observation of the weighted input pattern vector $\mathbf{o}$ is used as the *smoothed input pattern* for the Jordan network. Similarly, the tapped delay line at the output of the network provides a smoothed version of the classification result. The delay lines act as *weighted averaging windows* moving through time to get rid of random variations from the sequences.

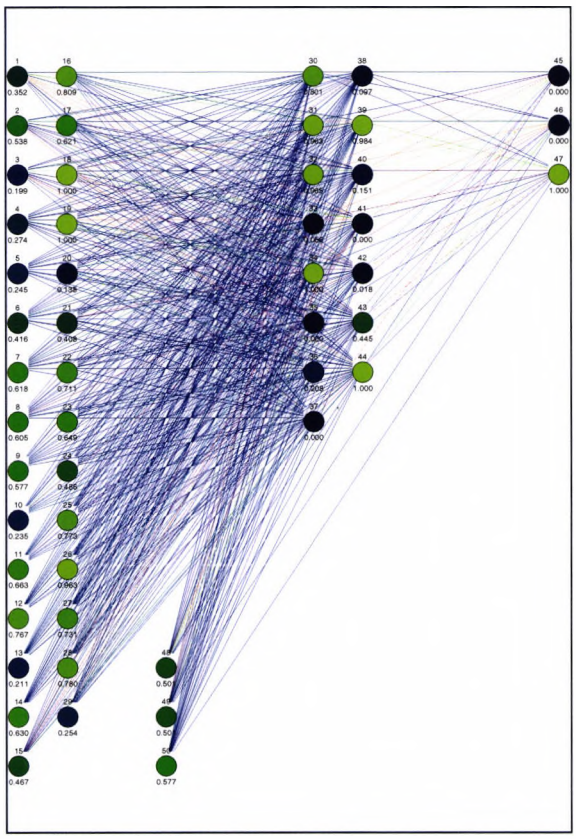
The proposed Jordan network is shown in Figure 5.8(a) while Figure 5.8(b) presents the overall structure of the classifier module. According to the assumption discussed in Section 5.3.1, the maximum time delay length of the proposed system is limited to 90 frames, allowing the system to monitor *three* seconds of the passenger history. Therefore, the sum of the lengths of the delay lines must satisfy the following condition:

$$(N+1) + (M+1) \leq 90 \text{ clocks}$$

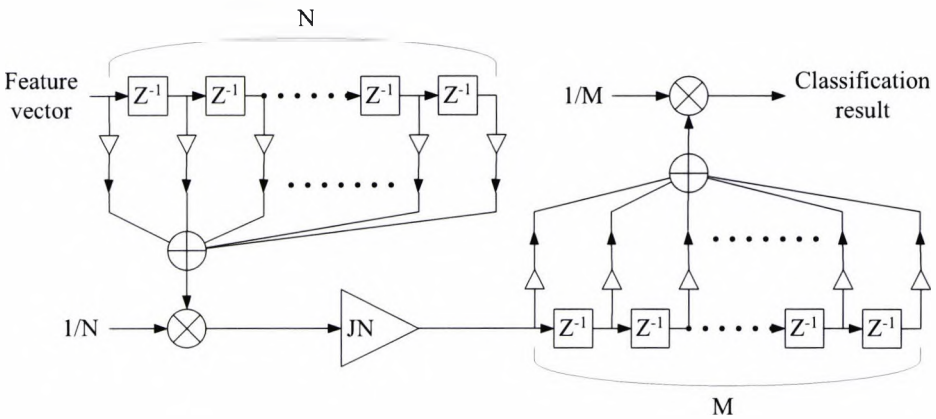
where N and M are the lengths of the *input* and *output* delay lines.

5.4 Precis

In this chapter, the design issue of a classifier for the occupant detection system is explored. The domain of the given classification problem is limited to solve a three-class problem since there is no need to detect empty seats as well as unknown objects for the safety reason. A number of the novel features to help effectively describe the tendencies of the occupant candidates in terms of two- and three-dimensional geometric aspects are proposed. These features are designed to distinguish at least one occupant class from the other two classes. A partially recurrent neural network is chosen as the classifier due to its ability to handle temporal data and the ease of network training. Two tapped delay lines with time-varying weights based on the motion information are employed to minimise the risk of over-concentrating on localised features and to reflect the dynamic characteristics caused by the movements of the occupants.



(a)



(b)

Figure 5.8: Classifier design: (a) the proposed Jordan network design after the learning process and (b) the proposed classifier framework with two tapped delay lines.

Chapter 6

Experimental results

6.1 Introduction 120
6.2 Experimental setup 120
6.2.1 Algorithm implementation and hardware embodiment 120
6.2.2 Data collection 122
6.3 Evaluation 123
6.3.1 Processing time 123
6.3.2 Feature consistency 124
6.3.3 Network training 124
6.3.4 Classification performance evaluation 125
6.4 Discussion 127

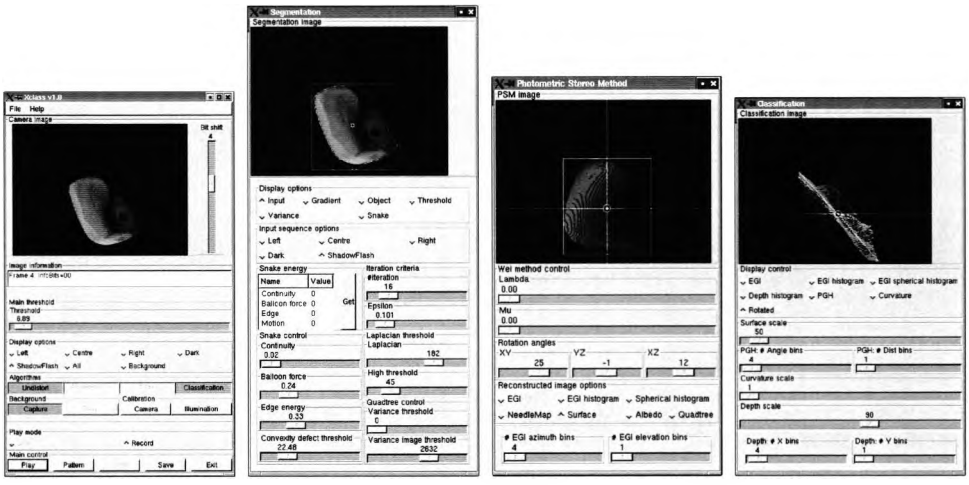


Figure 6.1: The proposed system implemented in the x-window environment.

6.1 Introduction

The aim of this chapter is to evaluate the applicability of the proposed framework discussed in the prior chapters to the given specific problem of classifying occupant types in a vehicle. The detailed description of the system setup for the experiments and the environmental conditions for data collection are discussed in Section 6.2. The quantitative evaluation of the proposed system according to the different types of the three occupant classes as well as the varying time periods for history observation are presented in Section 6.3. This is followed by a discussion on the overall system performance and analysis of the classification error cases in Section 6.4.

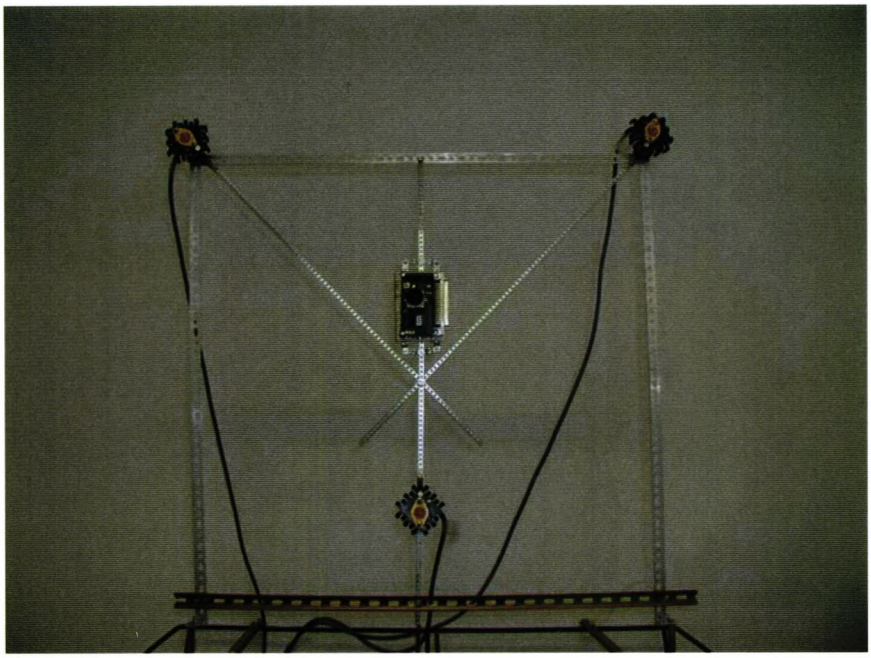
6.2 Experimental setup

6.2.1 Algorithm implementation and hardware embodiment

The algorithms are implemented in C/C++ programming languages with support of OpenCV libraries based on Linux environments [7]. A *multi-threaded* program is employed to minimise the risk of processing delay that might be caused by one of the algorithm modules. The advantage of using a thread group instead of a single serial program is the ability to carry out several operations in parallel. Thus, the overall system does not experience delay caused by the slowest algorithm module. GTK+ libraries are also utilised to provide the real-time GUI environment as shown in Figure 6.1.

The specification of the system used for the experiments is as follows:

- Pentium 4 processor at 2.8 GHz with 1024 MB RAMs



(a)

(b)

Figure 6.2: Experimental setup: (a) the illuminations are placed to form a right triangle around the imager, and (b) the setup simulates the situation where the imager is located at the roof console near the rear-view mirror in a vehicle cabin.

Figure 6.3: Typical samples of child seats.

- SCSI HDDs with the bandwidth of 160MB/s
- SollyCam with the dynamic range of 105dB (discussed in 2.2.2)
- Optical lens with the viewing angle of 120 degrees
- Digital frame grabber capable of 36-bit image data acquisition at 33MHz
- Three very high power infrared illuminators with peak emission at 880nm
- External controller for synchronising the infrared illuminators with the imager.

The arrangement of the imager and illuminations is shown in Figure 6.2(a). The three infrared illumination sources are placed around the imager forming a equilateral triangle, placing the imager in the center of the triangle. This geometry maximises both ShadowFlash and surface reconstruction performances. The location of the imager with respect to a passenger seat is determined as shown in Figure 6.2(b) to simulate the actual environment where the camera is located at the roof console of a vehicle cabin.

6.2.2 Data collection

The conditions under which these experiments were conducted were idealised in several aspects as compared to real-world conditions inside a vehicle. The sample sequences were collected in a laboratory environment which imitates a vehicle interior where the textures of the scene were simplified. 578 sample sequences were collected with a resolution of 320×240 in 12-bit gray scale at 30 Hz under varying illumination conditions provided by a number of high power halogen light sources. 199 sequences were recorded from 25 persons, while 379 sequences were taken from 29 different types of child seats which represented approximately 70% of available child seat products in the current European market. The length of the sequences varied from 100 to 500 frames depending on the occupant's behaviour pattern. The portions of the FFCS and RFCS classes in the child seat sequences made up roughly 50% each

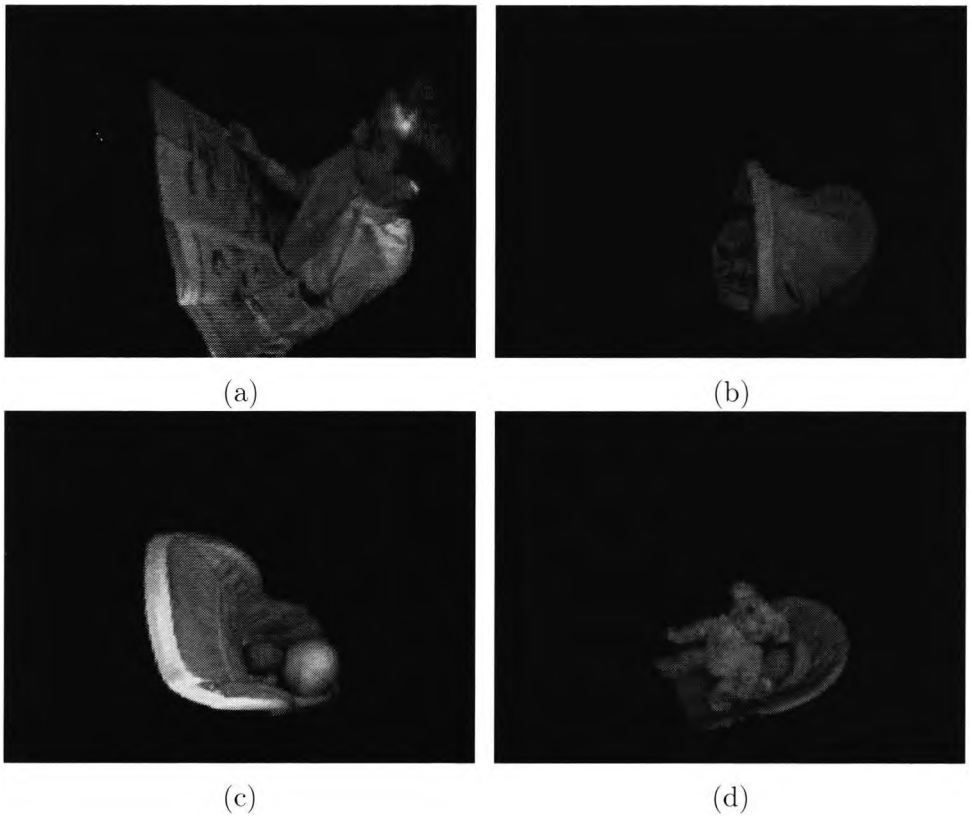


Figure 6.4: Various supplementary objects for providing the diversity of the test scene: (a) a passenger reading a newspaper, (b) a child seat covered by a blanket, (c) a child seat occupied by a baby holding a ball and (d) a teddy bear.

of the total sequences. The behaviour pattern of the passengers were not restricted to allow any possible movements in a vehicle setting. In order to simulate the diversity of the passengers in real environments, supplementary objects including some blankets, toys and newspapers were included in the sequence acquisition. Some typical examples of the child seats used in the experiments are shown in Figure 6.3, while Figure 6.4 presents examples of these additional objects employed in the experiments.

6.3 Evaluation

6.3.1 Processing time

Although the proposed system is designed to operate in real-time, the processing speed was not of great concern in evaluating the system as many improvements may be made by optimising the implementation prior to series production. In the experiments, an average time of 103.7 milliseconds for

	Pre-processing & segmentation	Surface reconstruction	Classification	Overall
Processing time (ms)	18.20 (17.55%)	53.44 (51.53%)	32.06 (30.92%)	103.7

Table 6.1: Processing time consumed in each processing module.

a single frame processing was achieved, mainly due to the supplementary codes for the GUI display in a non-embedded system environment. This result is approximately three times greater than the target processing time of 33 milliseconds (30Hz).

Table 6.1 shows the average processing time reserved by different processing stages. Over 50% of the total processing time was used for reconstructing object surfaces. This suggested that most effort should be focused on the optimisation of the 3D reconstruction algorithm module for improving the real-time performance of the system. Further improvements of the processing time is discussed in Section 7.4.

6.3.2 Feature consistency

A synthesised sequence with 30 input quadlets was used for evaluating the suitability of the proposed feature set. 10 of the 30 quadlets were randomly collected from each occupant class. The sequence was a collection of diverse scenes with varying illumination conditions, various types of child seats in different positions and directions, extreme motion and several additional objects used to diversify the scenes. The average values of some multi-dimensional features are used for simple display. The response of each feature detector is plotted in Figure 6.5. Although the responses of most proposed features were highly consistent within the same occupant class, some features such as the size of occupants, volumetric ratio, and compactness are relatively difficult to distinguish. These three features are strongly correlated with the volumetric sizes of vehicle occupants. Many earlier occupant detection systems implement volumetric attribute based object classification. However, experimental results show these features to provide insufficient distinguishability, which help to explain the difficulty of classification based on simple volumetric attributes of the target objects.

6.3.3 Network training

The sample sequences were divided into two groups creating a *training* and *testing set*, while the amount of sequences sampled from each class were evenly split in order to avoid unintended dominance of samples from any

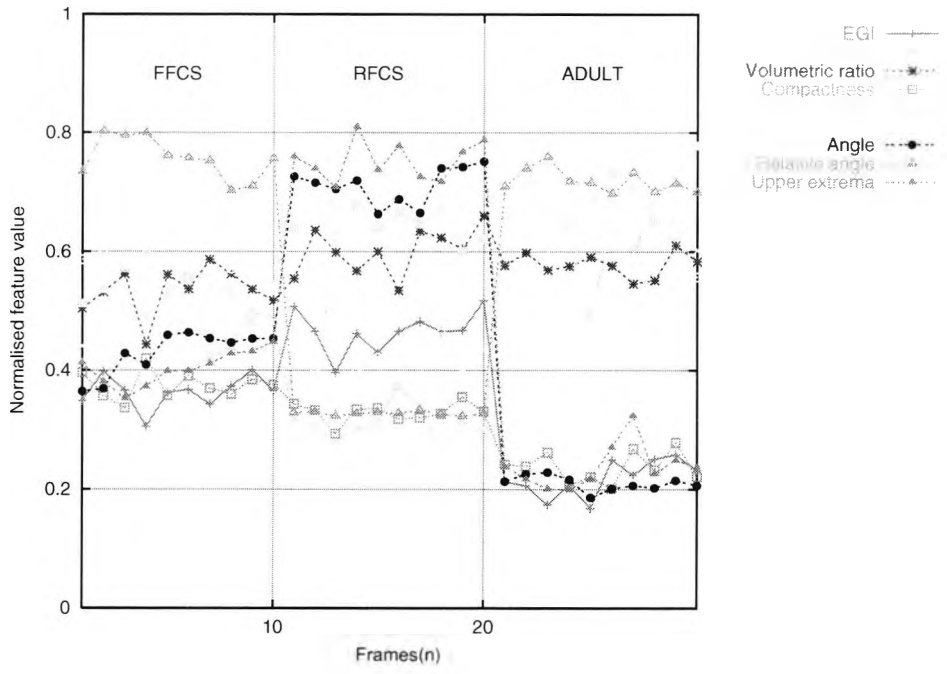


Figure 6.5: Feature consistency: the feature values with respect to the different types of occupant classes are compared.

one class. The patterns for the neural network were extracted from the sample sequences after the normalisation, while the actual occupant types were manually recorded.

The proposed Jordan network was trained by the resilient back-propagation (Rprop) algorithm with the *training* set [81]. The regular logistic activation function was set to all the neurons, and the initial values at the network's synapses were randomly chosen. In order to avoid overtraining the network, the learning was halted when the network reached the error minima where the mean squared output error reached 0.0793 after 120 iterations.

6.3.4 Classification performance evaluation

Since the neural network only makes a single frame decision, the classification performance was evaluated according to the lengths of two tapped delay lines using the *testing* set. The sum of the lengths of the delay lines was limited to under 90 frames. Two experiments were conducted for evaluating the effectiveness of employing the tapped delay lines and the motion-based weighting: (1) simple delay lines (no weights) and (2) delay lines employing the weights (weighted averaging).

Class type	FFCS	RFCS	Adult	Overall
Error rate(%)	14.2	15.4	0.725	6.66
Favourite error	RFCS(99.5%)	FFCS(90.0%)	RFCS(68.3%)	N/A

Table 6.2: Error statistics *without* the tapped delay lines. Overall error rate: 6.66%

Jordan network performance

Table 6.2 shows the error analysis according to the class types without support of the tapped delay lines. The overall classification error rate of the ordinary Jordan network reached 6.66%, which was comparable to the performances of the existing occupant detection systems based on vision technology discussed in Section 1.2.3.

Most errors occurred in discerning between the FFCS and RFCS classes due to their similar characteristics of the geometry caused by the alteration of additional objects. For example, a baby holding a teddy bear in the RFCS covered by a blanket coincidentally provided similar three-dimensional characteristics to the FFCS classes. Similar volumetric dimensions of child seats could be another factor which increases the ambiguity between the child seat classes.

However, low error rate in the adult class was achieved even with test sequences involving the large amount of motion. Most classification errors occurred in situations where the geometry of the target scene was significantly disturbed or occluded by additional objects (e.g. the passengers held unfolded newspapers in their movement). Contrast of clothing materials similar to the background was another source of errors causing false segmentation.

Network with non-weighted delay lines

The total number of the delay taps were limited to 90 to satisfy the *three seconds* requirement discussed in Section 5.3.4. No weights were employed for this experiment. Upon applying the tapped delay lines, the error rates of all classes were dramatically decreased as shown in Table 6.3. The best classification rate of 98.9% was achieved after setting the lengths of the input and output delay lines to 31 and 59, respectively.

In Figure 6.6, the classification error space computed by the proposed network with respect to the lengths of two delay lines is presented. It shows that the system is more sensitive to the length of the output delay buffer due to the recurrent network's adaptability to sequential behavior. However, as the sizes of both delay lines increases, the difference of the sensitivity becomes negligible.

Class type	FFCS	RFCS	Adult	Overall
Error rate(%)	10.1	13.7	0	1.14
Favourite error	RFCS(100.0%)	FFCS(91.7%)	N/A	N/A

Table 6.3: Error statistics *with* the tapped delay lines. Overall error rate: 1.14%

Class type	FFCS	RFCS	Adult	Overall
Error rate(%)	7.67	18.9	0.82	2.20
Favorite error	RFCS(59.0%)	FFCS(56.1%)	FFCS(73.2%)	N/A

Table 6.4: Error statistics of the system employing the weighted averaging approach for both input and output data streams. Overall error rate: 2.20%

Network with weighted delay lines

Another experiment employing the *weighted* delay lines were conducted using the Jordan network identically trained as the former experiment. The motion information was extracted from the acquired image sequences. The amount of motion was defined as the difference of the number of pixels which have intensity changes between a pair of adjacent frames. Table 6.4 shows the overall classification performance as well as the local error rates according to the different occupant classes when the motion based weighted averaging are applied to the delay lines. Although marginal improvement was expected from employing the weighted averaging scheme, results shows that the classification performance decreased by 1.1% compared to that of the network employing non-weighted delay lines. In the FFCS class, a performance improvement of 2.3% is gained, while no improvement is observed in the other two classes. In Figure 6.7, the plot of error surface with respect to the lengths of the delay lines is presented.

6.4 Discussion

Superior classification performance was achieved in the experiments compared to the existing vision based occupant detection systems introduced in Section 1.2.3. The performance of the proposed system in reality could be lower than the experimental results, considering that the sample sequences were collected under simplified conditions. Nevertheless, the goal of demonstrating the feasibility of an alternative classification system of comparable performance to binocular-based systems was successfully accomplished.

As most samples in the *adult* class had distinctive geometric dimensions compared to the two child seat classes, the lowest failure rate was achieved in this occupant type regardless of object motion. Most of the classifi-

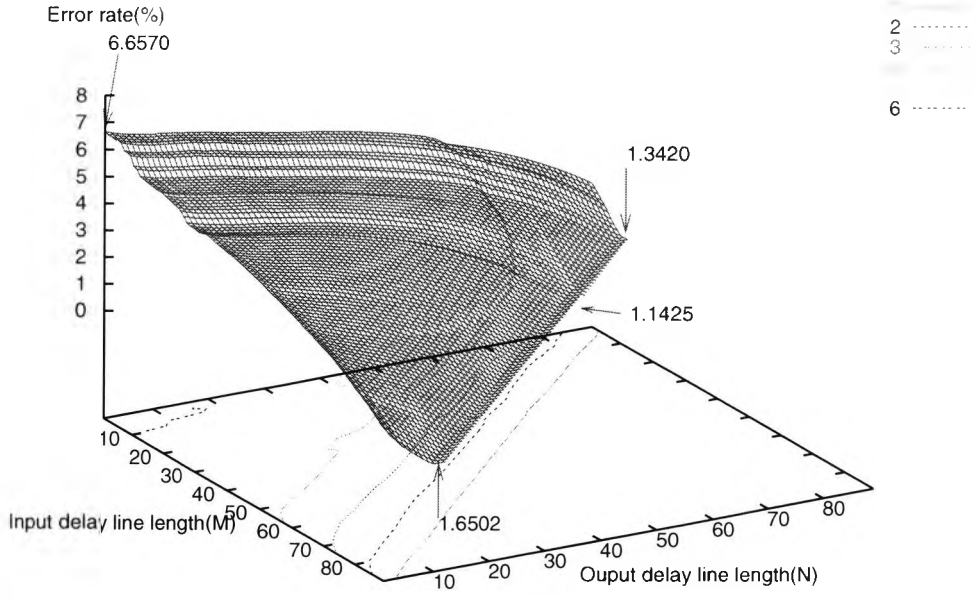


Figure 6.6: Classification error space produced by the proposed Jordan network with respect to different lengths of the tapped delay lines.

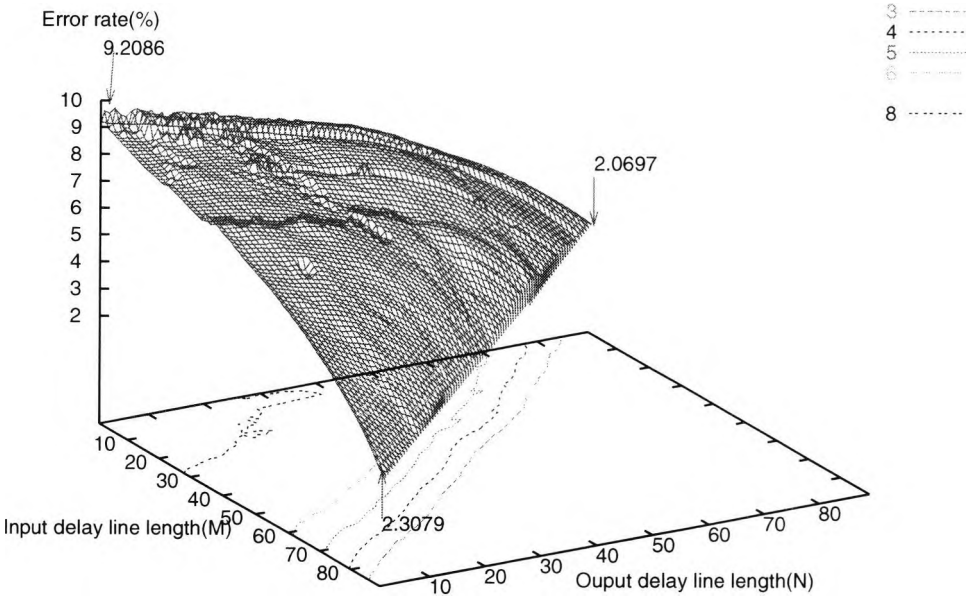


Figure 6.7: Error space plot of the classifier with the delay lines weighted by the motion information.

cation failures occurred between the FFCS and RFCS classes due to the compounded error factor from (1) the additional objects used to diversify the scenes, and (2) the false construction of the object surfaces caused by various noise sources discussed in Section 4.4.5.

Nevertheless, these are encouraging results, as the misclassification between an adult and child seat generally poses greater danger than that of the misclassification between two child seats. However, this suggests the necessity of developing new features for better discrimination of those child seat classes.

Two typical examples for misclassification are shown in Figure 6.8. In Figure 6.8(a), the passenger was recognised as a rear-facing child seat. The female in the scene was relatively small and comparable in size to some larger child seats. Some surface distortions were caused by the false segmentation result. Furthermore, the newspaper produced a secondary volumetric peak in the opposite side of the head position. This resulted in the erroneous estimation of the features associated with the spread axes as well as the depth property. The average surface normal direction used for estimating the EGI characteristics was also significantly affected by the normal vectors on the newspaper which generated strong directional tendency opposite to that of common adult types.

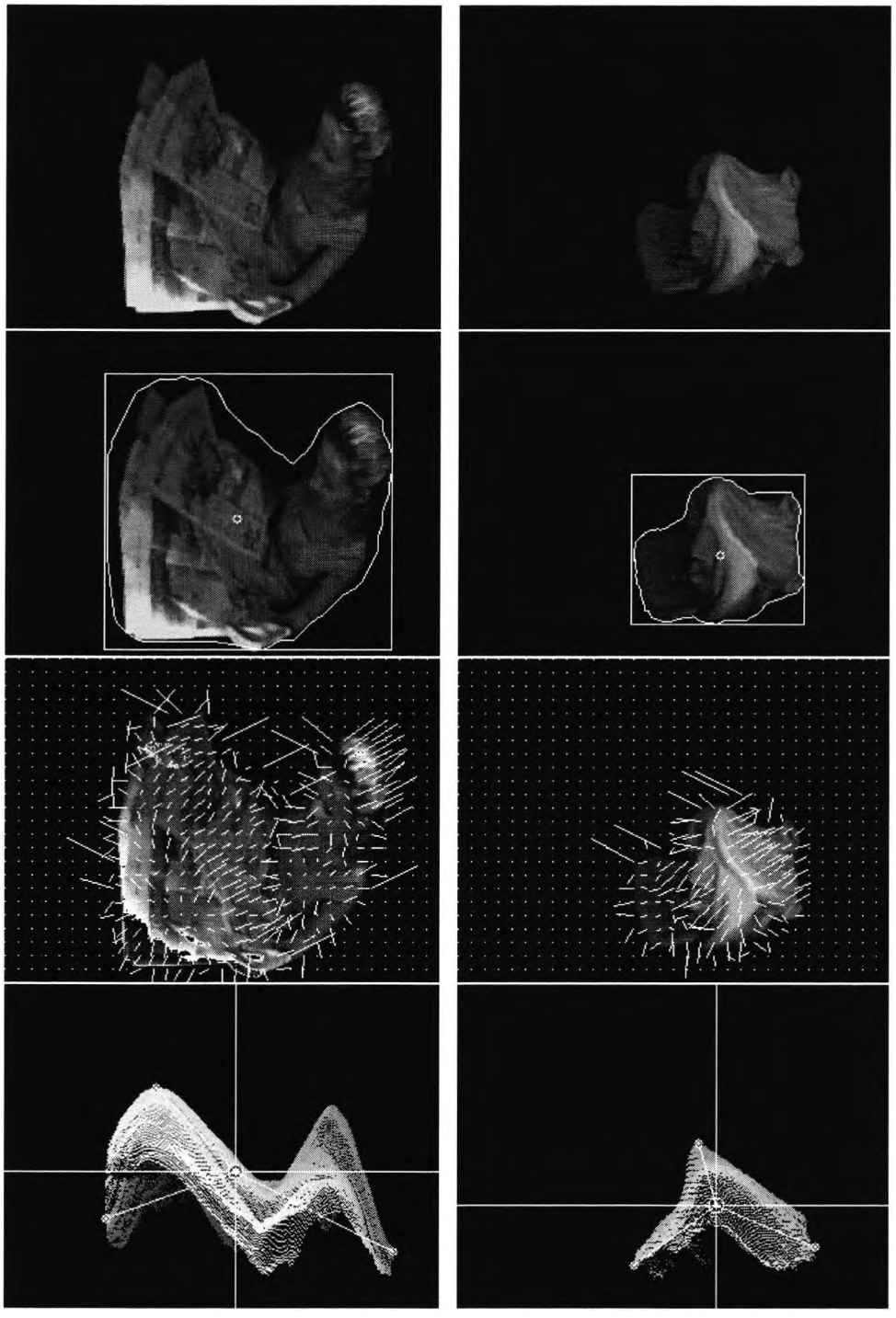
Figure 6.8(b) shows another misclassification example of a forward-facing child seat detected as a rear-facing child seat. The scene in this example simulates a situation where the carrying handle of the child seat is up and covered by a blanket in order to block direct sunlight from the baby. Although the segmentation as well as the estimation of the surface normals were successful, the limited capability of the photometric stereo method for a discontinuous object surface produced a smooth surface on the uncovered part of the child seat. The volumetric peak of the target object was located slightly closer to the front than the rear side of the child seat, which brought ambiguity to the estimation of both the depth histogram and the relative position of the upper extrema.

The maximum motion tolerance of the system was observed to be approximately between 5- to 10-pixel distance depending on the reflectivity of surface materials and the location of the point where the depth was estimated. For example, it was difficult to estimate the accurate surface normals using the brightness information taken from the points on surfaces with low and/or inhomogeneous reflectance such as human hair. The depth estimation of the point located on the object boundary was usually more sensitive to motion than the one in the middle of the surface with uniform reflectivity.

Another unsuccessful misclassification caused by severe motion is presented in Figure 6.9. The amount of motion involved in the original input sequence shown in Figure 6.9(a) exceeded the maximum tolerance of the

system, thus producing the significant distortions in the ShadowFlash results as shown in Figure 6.9(b). The severe motion resulted in the blurred ShadowFlash images, which caused the false estimation of the object boundary to include larger areas than the size of the actual object. Furthermore, the severe motion also brought a significant amount of unmatched areas in the PSM input triplets. In Figure 6.9(c), the inaccurately estimated surface normals were observed on the surface of the newspaper. Subsequently, the overall reconstruction process of the object surfaces was greatly hindered by the erroneous normal vectors due to the nature of the global minimisation used for the surface integration. Finally, the significantly distorted surfaces as shown in Figure 6.9(d) produced invalid feature vectors resulting in a misclassification.

The unstable response of the classifier in the temporal domain was successfully suppressed after employing the tapped delay lines capable of observing the feature history. The tolerance of the classifier in the temporal domain, however, may be further improved by employing alternative classifiers with more sophisticated memory structures such as *time-delay neural networks*. Unfortunately, the effectiveness of the *weighted* delay lines could not be proven in this work. There were two potential causes for the poor classification performances: (1) the motion information extracted by the relatively simple algorithm was not sufficiently correlated with the amount of the surface distortion; (2) some short sample sequences with limited motion were inappropriate for evaluating the system response with respect to the change of motion.



(a) (b)

Figure 6.8: Misclassification examples: (a) a passenger misclassified as a RFCS class, and (b) a forward facing child seat mistaken as a rear facing child seat. Each column shows images of the ShadowFlash result, the segmentation result, the needle map representation of the surface normals, and the recovered 3D surface of the target objects rotated by 90 degrees.

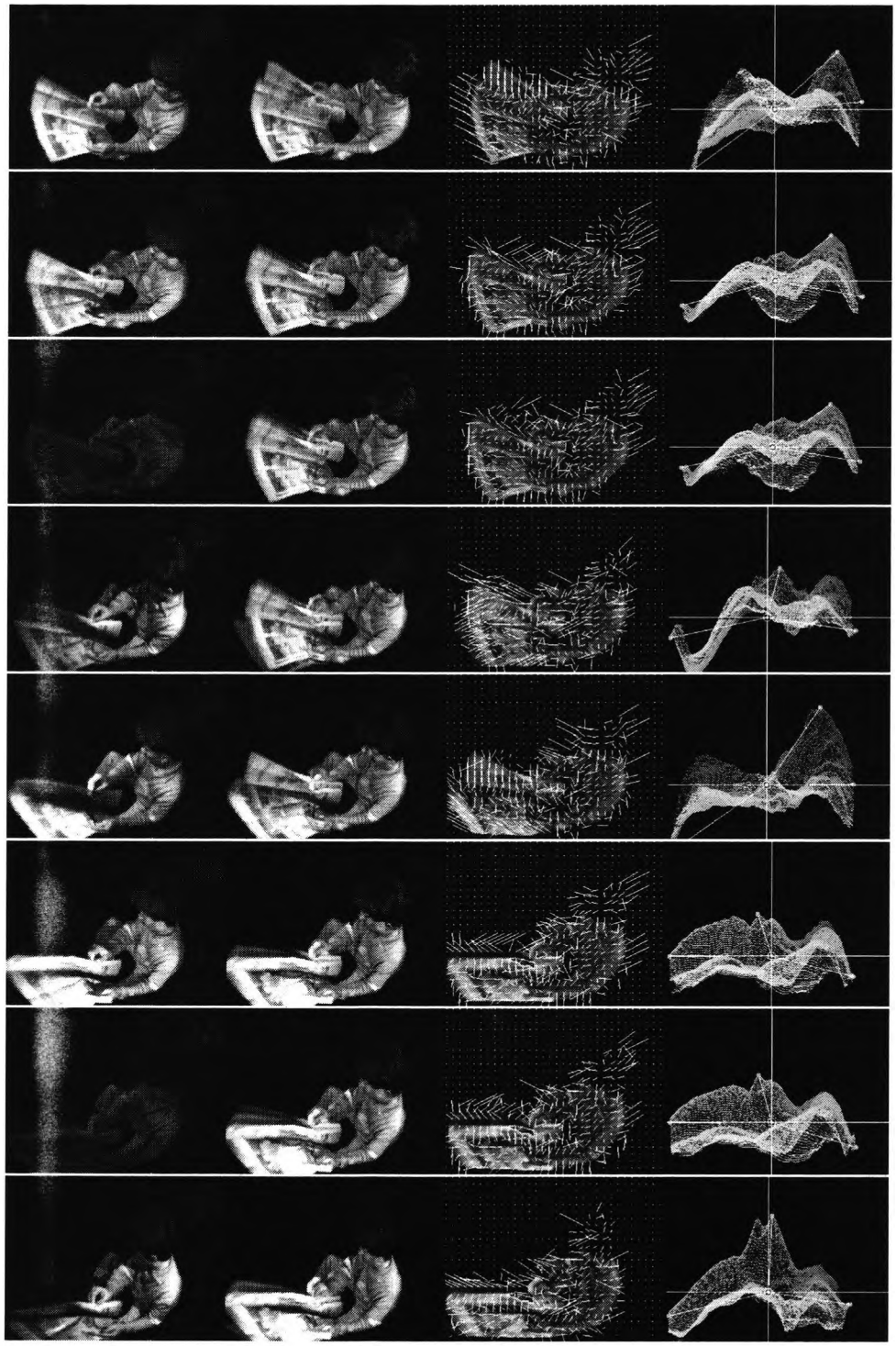


Figure 6.9: Sample sequence for the surface distortion caused by severe motion: (a) the original input sequence, (b) the ShadowFlash sequence, (c) the needle map representation of the estimated surface normals, and (d) the surface reconstruction results.

Chapter 7

Conclusions

7.1 Precis 134
7.2 Assessment 135
7.3 Contribution 136
7.4 Future work 137
Bibliography 139

7.1 Precis

In this thesis, a classification system based on three-dimensional information provided by a single camera with multiple illumination sources was proposed. The system is mainly designed to solve a problem of classifying vehicle passengers for safe airbag deployment. Most vision-based occupant detection systems developed in recent years suffered from vehicle environments characterised by large and frequent change of illumination conditions. Extreme contrast in a vehicle interior also required the high dynamic range of the imagers.

The proposed system was able to eliminate the influence of ambient illumination by employing the DoubleFlash technique originally introduced by Koch in [52]. This technique also compressed the dynamic range of the target scene by utilising two input images with different illumination power levels. The concept of the DoubleFlash technique was extended to the novel shadow removal technique, namely *ShadowFlash*. Cast shadows, frequently misclassified as imaginary objects, degenerated overall system performance. This technique suppressed such cast shadows based on the approach of simulating an infinite illumination plane by using a number of images illuminated by different light sources. By employing the *sliding N-tuple strategy*, the ShadowFlash method was extended to the *temporal domain* to provide shadowless scenes in real-time without reducing the original frame rate.

The extraction of the object boundary was an essential task to provide a priori knowledge about the two-dimensional geometry of the target object for subsequent processes. The similarity between an input frame and reference background was measured to create a set of target object candidates. The number of the candidates were reduced by the morphological operations which merged image blobs close in proximity. The image blob with the largest area would be designated as the target object. The boundary of the selected object was utilised as an initial boundary for active contour deformation. Dynamic programming was employed for efficient calculation of the optimal energy minimisation solution. The *concavity analysis* of the initial boundary provided the improved mobility to the active contour model.

The segmentation result was delivered to the surface recovery module for specifying the region of interest to reconstruct. To achieve the aim of designing a system capable of reconstructing three-dimensional object surface with minimal hardware costs in real-time, the photometric stereo method provided the optimal solution since the technique could exploit the existing illumination sources originally employed for the real-time ShadowFlash. After eliminating the influence of the ambient illumination from the input sequences using the DoubleFlash technique, three images subsequently captured under different illumination conditions were used to form an input *triplet* of the photometric stereo method. The technique calculated the sur-

face normal vectors by solving an albedo-independent irradiance equation which assumed a Lambertian surface. The integration of the object surface based on these normal vectors was performed by minimising several global constraints which controlled the integrability as well as the smoothness of the constructed object surface.

By ignoring non safety critical situations such as an empty seat, the given classification problem was simplified to a three-class problem. A feature vector with *29 dimensions* was defined based on the information extracted from both two- and three-dimensional geometry of the object of interest. The feature space was efficiently designed to discriminate at least one vehicle occupant class from the other two. A *partially-recurrent neural network* was selected as a classifier due to its superior ability to handle temporal sequences. To improve the dynamic property of the classifier, two *tapped delay lines* were employed to play the role of moving average windows.

The experiments were conducted under the idealised condition rather than a real vehicle environment. 578 passenger sequences including *adults*, *forward facing child seats*, and *rear facing child seats* were collected with a few additional objects to diversify the test scenes. Several supplementary illumination sources also provided various illumination conditions to the test sequences, emulating the possible illumination situations in a vehicle. The sequences were evenly split into the training and testing set to ensure no dominance by any occupant class. The learning process of the proposed neural network was performed with the training set. The classification rate of the partially-recurrent network was originally 93.3%. With the support of the tapped delay lines, the performance improved to 98.8%, with most misclassifications occurring between the FFCS and RFCS classes.

7.2 Assessment

This thesis proposed a novel structure of a real-time classification system operating in a high dynamic range environment. The main achievement of this work is the introduction of *multipurpose* active illuminations synchronised with a *single* imager. The purpose of active illumination can be summarised as

1. to provide the necessary light sources for the DoubleFlash technique to minimise the influence of ambient light fluctuations in a high dynamic range environment,
2. to help the ShadowFlash technique suppress unintended cast shadows without the distortion of texture details,
3. to provide the triple-active illumination required for the reconstruction of the three-dimensional surface of an object performed by the photometric stereo method.

By exploiting the existing illumination hardware for various purpose, low cost implementation and compact packaging of the system were accomplished.

The development of a novel *shadow removal approach* was another achievement in this work. The *ShadowFlash* technique simulated an infinite artificial illumination plane by combining multiple images captured under different illumination conditions, and successfully suppresses most undesirable shadows cast by strong light sources. Unlike other available shadow removal techniques, the proposed approach did not suffer from the degradation of textural details of target scenes. The simple algorithm caused *minimal processing overload* which allowed for implementations in low-cost real-time embedded systems.

This thesis also suggested a number of features designed for discerning vehicle occupants. The proposed features were (1) efficient to describe the geometric characteristics of the occupants, (2) robust to the interference caused by various noise sources, and (3) highly consistent in an occupant class. The proposed features could be utilised in any platforms for the extraction of both two- and three-dimensional information of objects.

7.3 Contribution

The main constraints to vision based passenger detection systems in the passenger vehicle environment are (1) cost, (2) system size, and (3) camera positioning. As the automotive sector is highly cost sensitive, any system must maintain sufficiently low cost to meet mass production requirement. Small camera size is advantageous for mounting in the limited space in the vehicle cabin. Furthermore, the space constraint dictates that only finite possibilities are available for placement of the camera system.

The two main paths to achieving successful occupant detection for the vehicle interior are (1) stereo vision based systems, and (2) monocular vision based systems. For a stereo camera system, it is more critical (compared to monocular systems) to have high flexibility for multi-camera placement in the limited cabin space. The monocular system addresses the limited cabin space problem with a different approach which inherently requires less space (e.g. using one camera instead of two).

The paramount issue in either approach is to provide techniques capable of accurate passenger classification under such tight constraints. The main contributions of this work can be divided into two areas: automotive safety applications, and machine vision systems. In the automotive arena, a major goal is to explore the feasibilities of the monocular approach in the current automotive technology and industry environment. It is also the hope of this work to further contribute to the field by introducing a novel monocular

camera occupant detection system as a stepping stone to future research and industry deployment. Finally, to the field of machine vision, the key contribution of this work is to extend the capability of the monocular vision system to cover applications requiring three-dimensional information in real-time.

7.4 Future work

Further research should be continued on the following topics:

ShadowFlash The proposed ShadowFlash technique is applicable to various environments experiencing frequent illumination changes such as a face recognition system of a cash dispenser or access authorisation system for building surveillance. The ShadowFlash method can significantly reduce the image processing cost and thus increase the recognition robustness if the systems have limited space, cost or design constraints.

Background maintenance As discussed in Section 3.2.1, the proposed framework was designed under the assumption of the fixed background model. For the realisation of the practical system, an improved background model capable of handling movable background objects such as a passenger seat should be developed. Alternatively, the core algorithms can be refined to make the system independent of the influence of any background change.

Moving object problem The proposed system experienced surface distortions caused by motion. The best way to overcome the moving object problem may be to simultaneously capture all the frames of an input quadlet. By employing multiple illumination sources with *different wave lengths* or a *multi-frequency spectrum light source* collaborated with corresponding optical bandpass filters, it should be possible to provide the input quadlets by splitting the captured image according to the wavelengths. An alternative solution for the moving object problem is to acquire an image by differently illuminating each *line* of the image sensor where the light sources are synchronised to the line clock rather of the sensor. The captured image is split into several sub-images, of which each image is influenced by a different light source [51].

Segmentation dependency In the proposed system, the quality of the reconstructed object surface is highly dependent on the segmentation result. A poorly-segmented object often involves areas which are not

appropriately flashed by the active illumination, and this distorts the recovered surface by producing inaccurate surface normals. The problem is especially serious if a global optimisation approach is employed for the surface integration. Therefore, the development/introduction of more effective boundary extraction techniques is an essential task to reduce the reconstruction sensitivity to the segmentation result.

Realistic reflection models The proposed system assumes an idealised condition where there is no inter- and intra-reflection in a target scene. However, in reality, the reflection properties are uncontrollable due to the various materials used in a vehicle interior and the complex reflections can interfere the surface normal estimation. Therefore, a realistic surface reflection model is necessary for improving system performance.

Classification Further improvement of the system performance could be achieved by various ways such as:

- introducing an alternative classifier more effectively handling time-series models (e.g. the idea of employing the tapped delay lines can be extended to employ alternative finite memory machines such as *time-delay neural networks*).
- utilising cluster analysis tools such as *principle component analysis* in order to minimise the dimensionality of the feature space.
- optimising the implemented algorithms for fast and efficient processing.
- providing an embedded platform with dedicated image processing units.

Out-of-position detection systems The next step of this work could be the extension of the proposed system to passenger out-of-position systems by providing the information about the precise position and pose of the occupants in real-time.

Other applications Finally, the proposed framework could be employed in various cost sensitive applications requiring non-contact three-dimensional imaging.

Bibliography

[1] A.A. Amini. Using Dynamic Programming for Solving Variational Problems in Vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(9):855–867, September 1990.

[2] M.S. Atkins and B.T. Mackiewich. Fully-automatic segmentation of the brain in mri. *IEEE Trans. on Medical Imaging*, 17(1):98–107, February 1998.

[3] S.S. Beauchemin and J.L. Barron. The Computation of Optical Flow. *ACM Computing Surveys*, 27(3):433–467, 1995.

[4] R.E. Bellman. *Adaptive Control Processes.* Princeton Univ. Press, Priceton, New Jersey, USA, 1961.

[5] M. Bichsel and A.P. Pentland. A simple algorithm for shape for shading. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 459–465, 1992.

[6] A. Blum. *Neural Networks in C++: An Object-Oriented Framework for Building Connectionist Systems.* John Wiley & Sons, 1992.

[7] G. Bradski. Opencv: Examples of use and new applications in stereo, recognition and tracking. In *Vision Interface 02*, page 347, 2002.

[8] K. Bubna and C.V. Stewart. Model selection techniques and merging rules for range data segmentation algorithms. *Computer Vision and Image Understanding*, 80:215–245, 2000.

[9] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. In *Proc. of the 5th IEEE International Conference on Computer Vision (ICCV)*, pages 694–699, Cambridge, MA, USA, 1995.

[10] A. Chakraborty, L. H. Staib, and J. S. Duncan. Deformable boundary finding influenced by region homogeneity. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 624–627, Seattle, WA, USA, June 1994.

[11] V. Chalana, W. Costa, and Y. Kim. Integrating region growing and edge detection using regularization. In *Proc. of the SPIE Conference on Medical Imaging*, 1995.

[12] C. Chesnaud, P. Réfrégier, and V. Boulet. Statistical region snake-based segmentation adapted to different physical noise models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(11):1145–1157, November 1999.

[13] A. Cleermans, D. Servan-Schreiber, and J.L. McClelland. Finite state automata and simple recurrent networks. *Neural Computation*, 1:372–381, 1989.

[14] L. D. Cohen and I. Cohen. Finite-element methods for active contour models and balloons for 2-d and 3-d images. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(11):1131–1147, November 1993.

[15] E.N. Coleman and R. Jain. Obtaining 3-dimensional shape of textured and specular surfaces using four-source photometry. *Computer Graphics Image Processing*, 18(4):309–328, April 1982.

[16] R. Cucchiara, C. Grana, M. Piccardi, and Prati A. Detecting moving objects, ghosts, and shadows in video streams. *IEEE Trans. on Pattern Recognition and Machine Intelligence*, 25(10):1337–1342, October 2003.

[17] C. Davatzikos and J.L. Prince. An active contour model for mapping the cortex. *IEEE Trans. on Medical Imaging*, 14:65–80, 1995.

[18] C. Davatzikos and J.L. Prince. Convexity analysis of active contour problems. *Image and Vision Computing*, 17(1):27–36, January 1999.

[19] J. Dayhoff. *Neural Network Architectures: An Introduction.* Van Nostrand Reinhold, NY, USA, 1990.

[20] A. Dempster, N. Laied, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(Series B):1–38, 1977.

[21] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern Classification.* Wiley, 2001.

[22] A. Elnagar and A. Basu. Motion detection using background constraints. *Pattern Recognition*, 28(10):1537–1554, October 1995.

[23] M.E. Farmer and A.K. Jain. Occupant classification system for automotive airbag suppression. In *Computer Vision and Pattern Recognition*, pages 756–761, Madison, Wisconsin, June 2003.

[24] D.A. Forsyth and J. Ponce. *Computer Vision: A Modern Approach.* Prentice Hall, 2003.

[25] E.R Fossum. CMOS image sensors: Electronic camea-on-a-chip. *IEEE Trans. on Electron Devices*, 44:1689–1698, October 1987.

[26] R.T. Frankot and R. Chellappa. A method for enforcing integrability in shape from shading problem. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 10(4):439–451, July 1988.

[27] P. Fua and C. Brechbuhler. Imposing hard constraints on soft snakes. In *Proc. of the European Conference Computer Vision'96*, pages 495–506, 1994.

[28] G. Funka-Lea and R. Bajcsy. Combining color and geometry for the active, visual recognition of shadows. In *the 5th International Conference on Computer Vision (ICCV95)*, pages 203–209, MIT, Cambridge, Massachusetts, USA, June 1995.

[29] D. Geiger, A. Gupta, L. Costa, and J. Vlontzos. Dynamic programming for detecting, tracking, and matching deformable contours. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(3):294–302, August 1993.

[30] R.C. Gonzalez and R.E. Woods. *Digital Image Processing.* Addison-Wesley Publishing Company, 1992.

[31] G.G. Gordon, T. Darrell, M. Harville, and J. Woodfill. Background estimation and removal based on range and color. In *Computer Vision and Pattern Recognition (CVPR99)*, pages II:459–464, June 1999.

[32] R.M. Haralick and L.G. Shapiro. *Computer Vision and Robot Vision*, volume I and II. Addison-Wesley, Reading, MA, USA, 1993.

[33] R.M. Haralick, S.R. Sternberg, and X. Zhuang. Image analysis using mathematical morphology. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 9:532–550, 1987.

[34] G. Healey and R. Jain. Depth recovery from surface normals. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 894–896, 1984.

[35] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

[36] B.K.P. Horn. Understanding image intensities. *Artificial Intelligence*, 8(2):201–231, April 1977.

[37] B.K.P. Horn. Height and gradient from shading. *International Journal of Computer Vision*, 5(1):37–76, August 1990.

[38] B.K.P. Horn and M.J. Brooks. Shape and source from shading. In *International Joint Conference on Artificial Intelligence*, pages 932–936, 1985.

[39] B.K.P. Horn and M.J. Brooks. The variational approach to shape from shading. *Computer Vision Graphics and Image Processing*, 33(2):174–208, February 1986.

[40] M.K. Hu. Visual pattern recognition by moment invariants. *IT*, 8(2):179–187, February 1962.

[41] J. Huppertz, R. Hauschild, B.J. Hosticka, T. Kneip, S. Müller, and M. Schwarz. Fast CMOS imaging with high dynamic range. In *IEEE Workshop on charge0coupled devices and advanced image Sensors*, pages R7:1–R7:4, Bruges, Belgium, June 1997.

[42] B. Jähne and H. Haußecker. *Computer Vision and Applications*. Academic press, 2000.

[43] A.K. Jain. *Fundamental of digital image processing*. Prentice Hall, 1989.

[44] M.I. Jordan. Attractor dynamics and parallelism in a connectionist sequential machine. In *Proc. of the 8th annual conference of cognitive science society*, pages 531–546, Amherst, MA, USA, 1986.

[45] B. Julesz and T. Caelli. On the limits of Fourier decompositions in visual texture preception. *Perception*, pages 8:69–73, 1979.

[46] M. Kass, A. Witkin, and D. Terzopolous. Snakes:Active Contour Models. In *International Conference on Computer Vision*, pages 259–268, 1987.

[47] K. Kinoshita and M. Lindenbaum. Cameral model selection based on geometric AIC. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 514–519, 2000.

[48] R. Klette, K. Schluens, and A. Koschan. *Computer Vision, Three-Dimensional Data from Images*. Springer, 1998.

[49] R. Klette and K. Schluüs. Height data from gradient fields. In *Photonics East'96*, SPIE Proceesings Vol.2908, pages 204–215, 1996.

[50] M. Klomark. Occupant Detection using Computer Vision. Master's thesis, Linköping University, SE-581 83 Linköping, Sweden, May 2000. LiTH-ISY-EX-3026.

[51] C. Koch. *Real-time occupant detection in hich dynamic range environments.* PhD thesis, City university, October 2003.

[52] C. Koch, S. Park, T.J. Ellis, and A. Georgiadis. Illumination technique for optical dynamic range compression and offset reduction. In *British Machine Vision Conference (BMVC01)*, pages 293–302, Manchester, England, UK, September 2001. BMVA Press.

[53] C. Koch, S.-B. Park, and S. Sauer. Method and apparatus for monitoring the interior space of a motor vehicle. International Patent EP 1.215.619, December 2000.

[54] D. Koller, K. Daniilidis, and H.-H. Nagel. Model-based object tracking in monocular image sequences of road traffic scenes. *Int'l Journal of Computer Vision*, 10(3):257–281, 1993.

[55] J. Krumm and G. Kirk. Video occupant detection for airbag deployment. In *IEEE Workshop on Applications of Computer Vision (WACV98)*, pages 30–35, Princeton, USA, October 1998.

[56] C.K. Lee and S.P. Wong. A mathematical morphological approach for segmenting heavily noise-corrupted images. *Pattern Recognition*, 29(8):1347–1358, August 1996.

[57] J.-M. Lequellec, F. Leasle, and S. Boverie. Car Cockpit 3D Reconstruction by a Structured Light Sensor. In *Intelligent Vehicles Symposium, 2000. IV 2000*, pages 87–92. Proceedings of the IEEE, October 2000.

[58] F. Leymarie and M. Levine. Tracking deformable objects in the plane using an active contour model. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(6):617–634, June 1993.

[59] W. Long and Y.H. Yang. Stationary background generation: An alternative to the difference of two images. *Pattern Recognition*, 23(12):1351–1359, 1990.

[60] R. Malladi, J. Sethian, and B. Vemuri. Shape modeling with front propagation: a level set approach. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 17(2):158–175, February 1995.

[61] S. Mann and R.W. Picard. On being 'undigital' with digital cameras: Extending dynamic range by combininb differenly exposed pixtures. *Society for Imaging Science and Technology*, (48):422–428, May 1995.

[62] A. Marín-Hernández and M. Devy. Application of a stereovision sensor for the occupant detection and classification in a car cockpit. In *International Symposium on Robotics and Automation*, pages 491–496, Monterrey, Mexico, November 2000.

[63] S. Maybank and P. Sturm. An assessment of information criteria for motion model selection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 47–52, 1997.

[64] I. Mikic, P. Cosmanand G. Kogut, and M. Trivedi. Moving shadow and object detection in traffic scenes. *International Conference on Pattern Recognition*, 1:321–324, September 2000.

[65] M.L. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry.* MIT Press, 1969.

[66] M.G.H. Mostafa, S.M. Yamany, and A.A. Farag. Integrating shape from shading and range data using neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages II: 15–20, 1999.

[67] N. Mukawa and H. Kuroda. Uncovered background prediction in interframe coding. *IEEE Trans. on Communications*, 33:1227–1231, 1985.

[68] National Semiconductor Corp. Application of the Piecewise Linear Response Feature in the LM9618/28 Image Sensors. Technical Report LM96198/28 application note 2, revision 1.1 edition, January 2002.

[69] W. Neuenschwander, P. Fua, G. Szekely, and O. Kubler. Making snakes converge from minimal initialization. In *Proc. of the 11th International Conference on Pattern Recognition*, pages 613–615, Jerusalem, Israel, 1996.

[70] R. Ohlander, K. Price, and R. Reddy. Picture segmentation by a recursive region splitting method. *Computer Graphics Image Processing*, 8:313–333, 1978.

[71] C.W. Omlin and C.L. Giles. Constructing deterministic finite-state automata in recurrent neural networks. *Journal of the Association for Computing Machinery*, pages 937–972, 1996.

[72] Y. Owechko, N. Srinivasa, S. Medasani, and R. Boscolo. Vision-based fusion system for smart airbag applications. In *IEEE Intelligent Vehicle Symposium*, Versailles, France, June 2002.

[73] S-.B. Park. *Optical vehicle compartment surveillance(in German).* PhD thesis, University of Duisburg, December 1999.

[74] S.-B. Park, A. Teuner, B.J. Hosticka, and G. Triftshaeuser. An interior compartment protecting system based on motion detection using CMOS imagers. In *International Conference on Intelligent Vehicles*, pages 297–301, October 1998.

[75] G. Paula. Sensors help make air bags safer. *Mechanical engineering magazine*, 119(8), 1997.

[76] N. Peterfreund. The Velocity Snake. In *Proc. of IEEE Nonrigid and Articulated Motion Workshop*, pages 69–79, june 1997.

[77] N. Peterfreund. Robust Tracking of Position and Velocity with Kalman Snakes. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(8), june 1999.

[78] I. Pitas and A.N. Venetsanopoulos. Morphological shape decomposition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 12(1):38–45, January 1990.

[79] R. Pless, J. Larson, S. Siebers, and B. Westover. Evaluation of local models of dynamic backgrounds. In *Computer Vision ans Pattern Recognition (CVPR03)*, pages II: 73–78, 2003.

[80] C. Ridder, O. Munkelt, and H. Kirchner. Adaptive Background Estimation and Foreground Detection using Kalman-Filtering. In *Proc. of International Conference on Recent Advances in Mechatronics (ICRAM95)*, pages 193–199, 1995.

[81] M. Riedmiller and H. Braun. A direct adaptive method for faster backpropagation learning: The rprop algorithm. In *Proc. of the IEEE International Conference on Neural Networks*, San Francisco, USA, 1993.

[82] M.A. Robertson, S. Borman, and R.L. Stevenson. Dynamic range improvement through multiple exposures. In *International Conference on Image Processing*, pages 159–163, Kobe, Japan, October 1999.

[83] R. Ronfard. Region-based strategies for active contour models. *International Journal Computer Vision*, 12(2):229–251, October 1994.

[84] P. Rosin and T. Ellis. Image difference threshold strategies and shadow detection. In *British Machine Vision Conference (BMVC95)*, pages 347–356. BMVA Press, 1995.

[85] D.E. Rumelhart, G.E. Hinton, and R.J. Williams. Learning representations by back-propagating errors. *Nature*, pages 533–536, 1986.

[86] E. Bienenstock S. Geman and R. Doursat. Neural networks and the bias / variance dilemma. *Neural Computation*, 4(1):1–58, 1992.

[87] S. Sarkar and K. Boyer. Quantivative measures of change based on feature organization: Eigenvalues and eigenvectors. *Computer Vision and Image Understanding*, 71(1):110–136, 1998.

[88] M. Seki, T. Wada, H. Fujiwara, and K. Sumi. Background subtraction based on cooccurrence of image variations. In *Computer Vision ans Pattern Recognition (CVPR03)*, pages II: 65–72, 2003.

[89] G.G. Sexton and X. Zhang. Suppression of shadows for improved object descrimination. In *IEEE Colloquium on Image Processing for Transport Applications*, pages 9/1–9/6, London, UK, December 1993.

[90] L.G. Shapiro, R.S. MacDonald, and S.R. Sternberg. Ordered structural shape matching with primitive extraction by mathematical morphology. *Pattern Recognition*, 20(1):75–90, 1987.

[91] D. Shen and C. Davatzikos. An adaptive-focus deformable model using statistical and geometric information. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):906–913, August 2000.

[92] J. Shi and J. Malik. Normalized cuts and image segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 731–737, 1997.

[93] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.

[94] C.G. Sodini and S.J. Decker. CMOS brightness adaptive imaging array with column-parallel digital output. In *IEEE Intelligent Vehicles Symposium*, pages 347–352, Stuttgart, Germany, October 1998.

[95] M. Sonka, V. Hlavac, and R. Boyle. *Image Processing, Ananlysis, and Machine Vision.* PWS Publishing, 1998.

[96] J. Stauder, R. Mech, and J. Ostermann. Detection of moving cast shadows for object segmentation. *IEEE Transactions on Multimedia*, 1(1):65–76, 1999.

[97] C. Stauffer and W.E.L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. of the Computer Vision and Pattern Recognition*, pages 23–25, 1999.

[98] K. Toyama, J. Krumm, B. Brumitt, and B. Meyers. Wallflower: Principles and Practice of Background Maintenance. In *Seventh International Conference on Computer Vision*, pages 255–261, 1999.

[99] National Highway Transportation and Safety Administration. Federal Motor Vehicle Safety Standard # 208. 2001.

[100] National Highway Transportation and Safety Administration (NHTSA). Occupant crash protection. Technical Report 49 CFR

Parts 552, 571, 585 and 595, Docket No. NHTSA 00-7013; Notice 1, RIN 2127-AG70, Department of Transportation (USA).

[101] M.M. Trivedi, S. Cheng, E. Childers, and S. Krotosky. Occupant posture analysis with stereo and thermal infrared video:algorithms and experimental evaluation. In *IEEE Transactions on Vehicular Technology, Special Issue on In-Vehicle Vision Systems.* Submitted August 2003.

[102] S.E. Umbaugh. *Computer Vision and Image Processing: A Practical Approach Using CVIPtools.* Prentice Hall, 1998.

[103] Y. Wang and L.H. Staib. Boundary finding with prior shape and smoothness models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(7):738–743, July 2000.

[104] T. Wei and R. Klette. Depth recovery from noisy gradient vector fields using regularization. In *Computer Analysis of Images and Patterns*, pages 116–123, 2003.

[105] Y. Weiss. Segmentation using eigenvectors: A unifying view. In *Proc. of the 7th International Conference on Computer Vision*, pages 975–982, 1999.

[106] D.J. Williams and M. Shah. A fast algorithm for active contours and curvature estimation. *Computer Vision Graphics and Image Processing*, 55(1):14–26, January 1992.

[107] A. Witkin, D. Terzopoulos, and M. Kass. Signal matching through scale space. *International Journal Computer Vision*, 1(2):133–144, 1987.

[108] H. Witters, T. Walschap, G. Vanstraelen, G. Chapinal, G. Meynants, and B. Dierickx. 1024×1280 pixel dual shutter APS for industrial vision. *SPIE Electronic Imaging*, 5017:19–23, January 2003.

[109] R.J. Woodham. Photometric method for determining surface orientation from multiple images. *OptEng*, 19(1):139–144, January 1980.

[110] P.L. Worthington and E.R. Hancock. New constraints on data-closeness and needle map consistency for shape-from-shading. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(12):1250–1267, December 1999.

[111] Z. Wu and R. Leathy. An optimal graph theoretic approach to data clustering: Theory and its application to image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 15(11):1101–1113, 1993.

[112] Z. Wu and L. Li. A line-integration based method for depth recovery from surface normals. *Computer Vision Graphics and Image Processing*, 43(1):53–66, July 1988.

[113] G. Xu, E. Segawa, and S. Tsuji. Robust active contours with insensitive parameters. *Pattern Recognition*, 27(7):879–884, 1994.

[114] D. Yang and A. El Gamal. Comparative analysis of SNR for image sensors with widened dynamic range. In *Proc. of the SPIE*, volume 3650, pages 22–28, San Jose, CA, USA, Feburary 1999.

[115] J. Yang and X.B. Li. Boundary detection using mathematical morphology. *Pattern Recognition Letters*, 16(12):1287–1296, December 1995.

[116] J.J. Yoon, C. Koch, and T.J. Ellis. Shadowflash: an approach for shadow removal in an active illumination environment. In *British Machine Vision Conference (BMVC02)*, pages 636–645, Cardiff, UK, August 2002. BMVA Press.

[117] R. Zhang, P.S. Tsai, J.E. Cryer, and M. Shah. Shape from shading: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 21(8):690–706, August 1999.

[118] Q. Zheng and R. Chellappa. Estimation of illuminant direction, albedo, and shape from shading. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 540–545, 1991.

[119] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust kalman filter. In *International Conference on Computer Vision (ICCV03)*, pages 44–50, October 2003.

[120] Y. Zhou and H. Tao. A background layer model for object tracking through occlusion. In *International Conference on Computer Vision (ICCV03)*, pages 1079–1085, October 2003.

[121] S. C. Zhu, T. S. Lee, and A. L. Yuille. Region competition: unifying snakes, region growing, energy/bayes/mdl for multi-band image segmentation. In *Proc. of the 5th IEEE International Conference on Computer Vision*, pages 416–423, Cambridge, MA, USA, 1995.

Index

active contour model, 48, 49, 53
  external energy, 57
  internal energy, 55
    balloon force, 55, 56
    continuity energy, 55
airbag, 3
  advanced, 5
  de-powered, 6
  multi-stage, 6
  smart, 5
albedo, 84, 85
automatic gain control, 91

background
  fixed, 51
  maintenance, 50
  model, 50
  reference, 49
bandwidth, 74
BMW, XXI
boundary
  approximate, 49
  initial, 49
boundary extraction, 29

camera
  binocular, 9
  CCD-based, 2, 9, 18
  CMOS-based, 2, 8
    active pixel sensor, 20
    advantages of, 20
    low power consumption, 21, 23
    system-on-chip integrability, 21
  monocular, 9
  MOS-based, 18
  multi-, 10
  multiple, 81
  response characteristic, 21
    linear, 21
    non-linear, 21
    piecewise linear, 22, 23
  single, 10
  SollyCam, 23, 122
  stationary, 29
  the number of, 10
  TIR, 9
  trigger clock, 11
  trinocular, 9
classifier, 102
  k-nearest neighbour, 8, 103
clustering, 47
compactness, 108
computational complexity, 90
concavity analysis, 49
console box, 82
convex hull, 49, 54, 60
convexity defects, 49, 54, 60, 61
correspondence analysis, 80
  feature-based, 80
  intensity-based, 80, 81
curse of dimensionality, 104

delay
  frame, 105
  inter-frame, 91
depth
  absolute, 89
  relative, 83, 89, 93
depth map, 80, 81
depth resolution, 82
depth step, 82
discontinuity, 46

disparity, 80, 81
DoubleFlash, 11, 16, 26, 134
    dynamic range compression, 27
    offset reduction, 26, 36
downsampling, 91
dynamic programming, 57

expectation maximisation, 47
extended Gaussian image, 104

fast Fourier transform, 89
feature, 102
    distinguishing, 104
    selection, 104
    space, 102, 109, 135
    variability, 102
    vector, 11, 102
fixed pattern noise, 51
focal length, 82
Fourier basis function, 89
frame rate
    sufficient, 91
Frankot-Chellappa algorithm, 84
fuzzy logic, 103

gamma correction, 91
generalisation, 111
gradient vector, 84
gray code, 11
GTK+ libraries, 120

Hu moments, 108

illumination
    active, 24, 81
    ambient, 11, 29, 50
    condition, 50
    direction, 84, 85
    infrared, 11, 24
    multiple, 83
    passive, 24
illumination direction vector, 86
infinite illumination plane, 31, 36, 41, 91
input quadlet, 62, 92
input triplet, 92, 134
integrability, 84, 88
irradiance equation, 85

Lambert's cosine law, 84
Lambertian surface, 84, 85
Legendre moments, 103

morphology, 52
    closing, 52
    dilation, 52
    erosion, 52
    opening, 52
    structuring element, 52, 53
motion
    detection, 91
multi-threaded program, 120

neighbourhood matrix, 58
neural network, 11, 110
    activation function, 111
    artificial neuron, 110
    back propagation, 114
    bias, 111
    connection weight, 110
    feed-forward, 112
    Jordan network, 114
    perceptron, 112
    processor, 110
    recurrent, 112, 113
        fully, 113
        partially, 104, 114
    testing, 124
    training, 124
NHTSA, 4
    FMVSS 208, 5
non-rigid object, 48, 60, 81
normalised central moments, 108

occupant class, 6, 11
    FFCS, 6
    NOPS, 6
    ODFC, 6
    PCSP, 6
    POOP, 6

RFCS, 6
occupant detection system
load sensor, 7
manual switches, 7
transponder, 7
vision-based, 3, 8
multiple camera approaches, 9
single camera approaches, 8
optical dynamic range, 17, 89
global, 17
high dynamic range environment, 18, 20
high dynamic range environments, 3
local, 17
of CCD cameras, 2
optical flow, 91
optical lens distortion, 87
out-of-position, 5
detection system, 6
out-of-position detection system, 11

photometric stereo method, 11, 49, 83, 85
albedo-independent, 87
pre-processing, 16
principle component analysis, 138
processing power, 74

quadtree subregioning, 105

radiance equation, 84
reflectance map, 84
Lambertian, 84
reflectivity, 27
rotational transformation, 105

segmentation, 46
elastic models, 47
probabilistic models, 47
region-based approaches, 46
self-reflection, 91
sensor noise, 87
ShadowFlash, 11, 16, 83, 134
real-time, 37
shadows, 3, 28, 29
penumbra, 30
umbra, 30
shape from shading, 83, 85
similarity, 46, 81
sliding N-tuple strategy, 29, 37, 41, 134
snake, 48, 53
snaxel, 54
spread angles, 103, 105
relative, 105
spread axes, 105
statistics
first-order, 51
second-order, 51
subimaging, 51
subpixel, 81, 83
surface integration
global minimisation, 88
local propagation, 88
surface normal, 84, 86–88, 90

tapped delay line, 11, 104, 114, 115
weighted, 115, 127, 130
template matching, 102
texture
analysis, 51
primitive, 51
three-dimensional
information, 74, 87
sensing techniques, 74
digital photogrammetry, 75
laser scanning, 75
shape-from-shading, 76
stereo vision, 77, 78
structured lighting, 9, 75, 76
time-of-flight, 75, 76
time-of-flight measurement, 75
triangulation, 75
ultrasonography, 75
surface reconstruction, 74
time-delay neural network, 130, 138

uniqueness theory of moments, 108

vehicle interior monitoring system, 90
vibration, 82
volumetric
    density, 103
    ratio, 108