3D-Rigid Objects Motion Segmentation: A Study of Practical Limitations

Shafriza Nisha Basah

Submitted in total fulfilment of

the requirement for the degree of

Doctor of Philosophy

Faculty of Engineering and Industrial Sciences

Swinburne University of Technology

June 2011

Abstract

Motion segmentation or recovering structure-and-motion (SaM) from images of dynamic scenes plays a significant role in many computer-vision applications ranging from local navigation of a mobile robot to image rendering for multimedia applications. Since in many applications, the exact type of motion and camera parameters are not known, *a priori*, the fundamental matrix is commonly used as a general motion model. Although the estimation of a fundamental matrix and its use for motion segmentation are well understood, the studies of conditions governing the feasibility of segmentation for different types of motions are largely unaddressed.

In this thesis, the feasibility of motion segmentation using the fundamental matrix is analysed. The focus is on a scene including multiple SaMs viewed by an uncalibrated camera. The quantifiable measures for the degree of separation were theoretically derived for the types of motion that are usually seen in practical applications, namely, motion from background, translational motions and planar motions. Sets of condition to guarantee successful segmentation were proposed via extensive experiments, the design of which was based on the Monte Carlo statistical method, using synthetic images. Experiments using real image data were set up and executed to examine the relevance of those conditions to the problems encountered in real applications.

The experimental results show the capability of the proposed conditions to correctly predict the outcome of several segmentation scenarios. In addition, they also show that the Monte Carlo experimental results are very relevant to the problems encountered in real applications. In practice, the success of motion segmentation could

be predicted via the value of the degree of separation between two motions estimated from obtainable scene and motion parameters. Therefore, the proposed conditions serve as a guideline for practitioners in designing motion segmentation solutions for computer-vision applications.

Acknowledgements

First and foremost, I offer my utmost gratitude to Allah, the most gracious, the the most merciful, for giving me and my family good health, mental and physical strength, and importantly, an uncomplicated life over the years.

I would like to extend my sincerest appreciation to my supervisor, A.Prof Alireza Bab-Hadiashar, who, with his knowledge and wisdom, has provided precious advice, support and encouragement over the years. It was he who introduced me to the potentially new research directions in the exciting areas of motion segmentation and robust estimation, and who made a thorough proof-reading, providing valuable input to improve the clarity of the thesis and academic papers. In addition, he was always approachable and offered unfailing assistance. As a result, research life was smooth and became a fulfilling experience for me.

I would like to thank my associate supervisor — Dr. Reza Hoseinnezhad who equipped me with the necessary fundamental theories and a clear picture of the research background. His important contributions at the early stage of my study managed to steepen my learning curve, ensuring a quick transition from the literature review to achieving my project goals. Additionally, Dr. Reza's enthusiasm for research and his highly motivational personality have been very inspiring.

I appreciate the technical discussions with Dr. Niloofar Gheissari and Dr. Reyhaneh Hesami about the fundamental-matrix estimation, robust estimator and its implementation using MATLAB. I thank Dr. Jean-Yves Bouguet, Prof. David Lowe and Dr. Peter Kovesi for their publicly available codes for camera calibration, feature extraction and fundamental-matrix estimation. These codes have assisted in

developing the experiments using synthetic images as part of this project.

In the laboratories and workshops, I have been aided by a number of very capable technicians — Mr. Walter Chetcuti, who ensured the camera peripherals were in working order, Ms. Meredith Jewson and Mr. David Vass, who accurately fabricated all three-dimensional models for the experimental purposes. In addition, many thanks to Ms. Gin Tan from the Information Technology Services (ITS), Swinburne University, Dr. Jiro Doke from MathWorks, and Mr. Lev Lafayette and Ms. Jin Zhang from the Victorian Partnership for Advanced Computing (VPAC), who tirelessly spent hours in setting up, debugging and testing to ensure that my experiments using a supercomputer cluster in VPAC could be smoothly executed without interruption.

I am also indebted to many anonymous reviewers of my journal and conference papers. Their valuable reviews and comments have indeed improved the clarity of the ideas in each of my papers. My appreciation also goes to Ms. Ruth Fluhr, who meticulously edited the draft of this thesis and provided suggestions which have been used to improve the presentation of the thesis.

Last but not least, my gratitude goes to my family — to my parents, who give me their prayers and solicitude, and to my loving wife, who sacrificed her career to take care of me and our growing children, so that I could give my undivided attention to completing my PhD. During the course of this thesis, we were blessed with our third *bundle of joy* — a baby girl, named Aiesyah.

For my parents,

my wife, Ana,

&

our children,
Khadijah,
Hamzah,
&
Aiesyah.

Declaration

I certify that this thesis is the result of my own research and to the best of my knowledge, this thesis contains no material published elsewhere except where due reference is made and acknowledged. This thesis, in whole or in part, has not been submitted previously for the award of any degree in any other tertiary institution.

Shafriza Nisha Basah

June 2011

Melbourne, Australia

Publications

Parts of this thesis have been published or are currently under review/preparation for publication:

Published and accepted papers

1. S. N. Basah, A. Bab-Hadiashar, and R. Hoseinnezhad. Conditions for motion-background segmentation using fundamental matrix. *IET Computer Vision*, 3(4):189-200, 2009.

2. S. N. Basah, R. Hoseinnezhad, and A. Bab-Hadiashar. Limits of motion-background segmentation using fundamental matrix estimation. In *Digital Image Computing: Techniques and Applications DICTA '08*, pages 250-256, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

3. S. N. Basah, A. Bab-Hadiashar, and R. Hoseinnezhad. Conditions for segmentation of 2d translations of 3d objects. In *Proceedings of the 15th International Conference on Image Analysis and Processing ICIAP 09*, volume 5716 LNCS, pages 82-91, Berlin, Heidelberg, 2009. Springer-Verlag.

4. S. N. Basah, R. Hoseinnezhad, and A. Bab-Hadiashar. Conditions for segmentation of motion with affine fundamental matrix. In *Proceedings of the 5th International Symposium on Advances in Visual Computing ISVC 09*: Part I, volume 5875 LNCS, pages 415-424, Berlin, Heidelberg, 2009. Springer-Verlag.

Papers under review or in preparation

1. S. N. Basah, A. Bab-Hadiashar, and R. Hoseinnezhad. Analysis of Planar-Motion Segmentation Using Fundamental Matrix. *Under review for IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics.*

2. S. N. Basah, A. Bab-Hadiashar, and R. Hoseinnezhad. Analysis of Translational Motion Segmentation Using Fundamental Matrix. *In preparation for Image and Vision Computing.*

Contents

List of Tables

List of Figures

Chapter 1

Introduction

This chapter presents an introduction to the research work in this thesis. The background and motivation of the research work are explained to establish the research question. In addition, the overview of the research aim, objectives and contributions is described. Finally, the chapter provides the structure of the thesis.

1.1 Background and motivation

Motion segmentation aims to recover structure-and-motion (SaM) from a collection of two-dimensional (2D) camera images of a dynamic scene. In practice, the motion-segmentation problem can be far more challenging, especially in situations where multiple objects having different motions are present in the scene (called multibody structure-and-motion (MSaM) by Schindler and Suter [80]). It is an important problem as it forms the initial step in many computer-vision applications, such as in robotics, traffic and video surveillance, assembly inspection, object recognition, local navigation, image rendering and many others. The interest in SaM or MSaM

recovery from multiple views stems from the established research in the recovery of object shapes, summarised in [25, 39, 61], or traditionally termed *Structure from Motion* (SfM). SfM considers images of static objects viewed by a moving camera [26, 38] whereas MSaM recovery deals with a dynamic scene including multiple objects having distinct motions [80, 81].

Generally, motion segmentation is a complex process which involves three main tasks — feature extraction, motion modelling and segmentation strategy [123]. The main problem in SaM or MSaM recovery is that the exact nature of objects' motions and the camera parameters are often not known in advance; thus, the most general motion model in the form of a fundamental matrix is preferred to model a three-dimensional (3D) rigid-moving object [104]. The fundamental matrix encapsulates the geometry of a 3D structure, its motion and the camera intrinsic parameters [27, 33, 123]. In addition, using the fundamental matrix could eliminate the need for camera calibration, which is advantageous for demanding applications where the camera parameters can be constantly changing due to zooming effect or external vibration [14, 100].

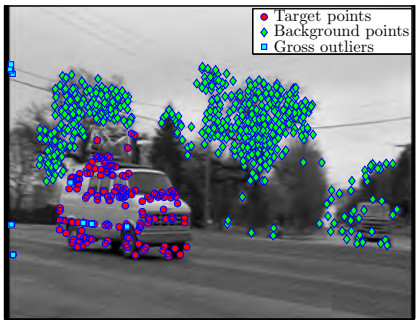
A number of techniques for the estimation of multiple fundamental matrices, fundamental matrix approximations and their use in motion segmentation have already appeared in the computer-vision literature and are summarised in [25, 39, 61]. However, the conditions governing the feasibility of motion segmentation of each SaM are yet to be established.

In order to demonstrate a motion-segmentation process, its challenges and limitations, two simple examples are presented. In these examples, we have performed motion segmentation to recover the light-coloured van from images in figures 1.1(a) and

1.1(b) [78] and the object on the left hand side of images in figures 1.2(a) and 1.2(b) [78], respectively. All corresponding feature points in both images in figures 1.1(a)-1.1(b) and 1.2(a)-1.2(b) were extracted using the publicly available implementation of the Scale-Invariant Feature Transform (SIFT) algorithm [58, 56]. The ground truth of the extracted points, as shown in figures 1.1(c) and 1.2(c), are associated with

(a) Image-1

(b) Image-2



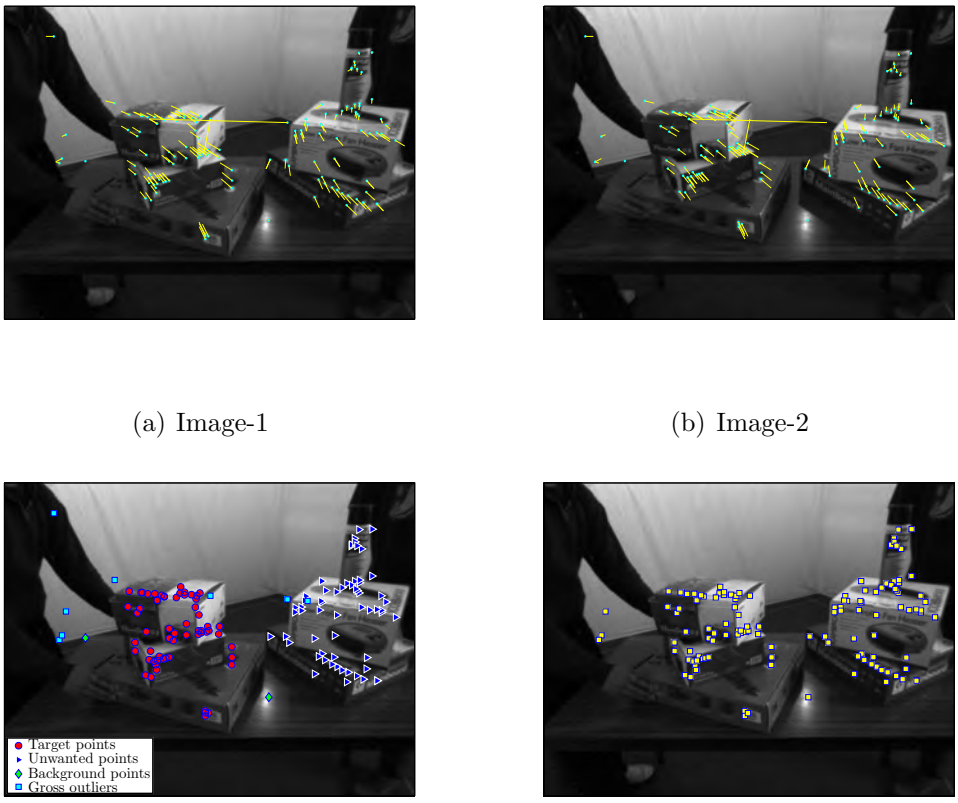
(c) Ground-truth

(d) Segmentation result

Figure 1.1: Corresponding image points from two images of a 3D object in motion, in (a) and (b), with the light-coloured van being the target object. The ground-truth in (c) and the segmentation results are superimposed in image-1 in (d).

each moving object, the background and some of them are the mismatches or *gross outliers*.

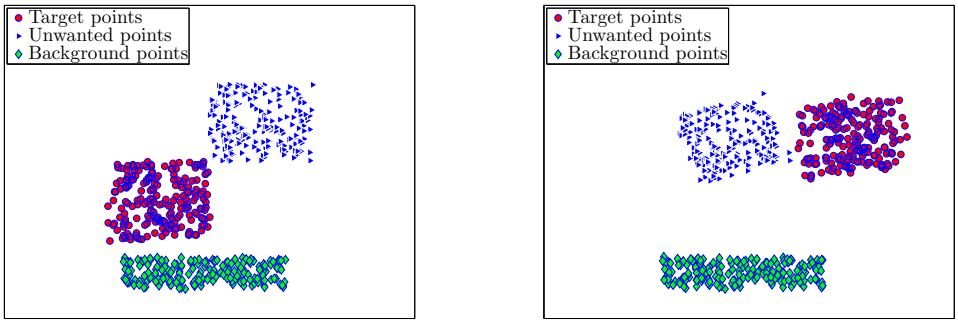
The fundamental matrix associated with the target object and motion for each case, shown in figures 1.1 and 1.2, was estimated using a publicly available function for the normalised *eight-point algorithm* [39, 52] with random sampling and



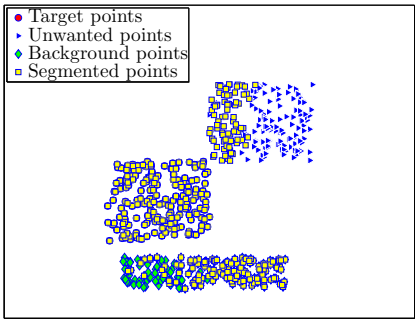
(a) Image-1 (b) Image-2

(c) Ground-truth (d) Segmentation result

Figure 1.2: Corresponding image points from two images of two 3D objects in motion, in (a) and (b), with the object on the left being the target object. The ground-truth in (c) and the segmentation results are superimposed in image-1 in (d).



(a) Image-1 (b) Image-2



(c) Segmentation result

Figure 1.3: Corresponding image points from two images of a simulated dynamic scene, in (a) and (b). The segmentation results using the true fundamental matrix are superimposed in image-1 in (c).

motion segmentation was concurrently performed using the Random Sample Consensus (RANSAC) [28] as a gold-standard robust estimator having taken the square of Sampson distances [101, 104, 119] as the residuals[1]. Figures 1.1(d) and 1.2(d) show that the segmentation results are wrong as a number of points from the background

[1]Details on the fundamental matrix, its estimation, robust estimator and Sampson distance are covered in chapter 2.

or the other object are segmented as the target object. The potential sources of error that lead to the unsuccessful segmentations include incorrect estimation of the fundamental matrix, the presence of *gross outliers* in the data and/or the limitation of the fundamental matrix motion model.

To investigate the dominant source of segmentation errors in figures 1.1(d) and 1.2(d), we have simulated a dynamic scene containing two 3D objects having random motions and a static object representing the background. In the simulation, all feature points associated with both moving and static objects were projected onto two 512×512 images and the ground-truth feature points are shown in figures 1.3(a) and 1.3(b). The measurement noise in the image data was assumed to be Gaussian distribution with standard deviation around one pixel. The camera was calibrated and the scene was controlled such that the true fundamental matrix of the target motion was known and no *gross outliers* were present in the data.

Motion segmentation was performed to segment the points associated with the target motion using its true fundamental matrix. As shown in figure 1.3(c), the segmentation is wrong even though the true fundamental matrix was used as the motion model and the segmentation was performed without any *gross outliers*.

The segmentation results in figures 1.1(d), 1.2(d) and 1.3(c) imply that the success of motion segmentation depends on the motion and/or scene parameters. The research question deduced from the preliminary analysis is: *What type and how much motion can be segmented using the fundamental matrix?* To answer this question, a feasibility analysis for successful motion segmentation needs to be thoroughly conducted to determine the theoretical limits for correct and successful segmentation using the fundamental matrix motion model. The governing conditions for successful

segmentation can be developed from those theoretical limits. In practice, these conditions are capable of predicting the outcome of motion segmentation in advance and they provide a useful set of guidelines for practitioners designing motion-segmentation solutions for computer-vision applications.

1.2 Aim and contributions

The thesis aims to analyse the feasibility of motion segmentation using the fundamental matrix motion model. The focus is on a dynamic scene involving multiple rigid 3D objects viewed by an uncalibrated camera. The analysis is divided into three main parts based on the types of motions that are commonly encountered in computer-vision applications. These parts are: stationary object from the background, translational motion and planar motion. The objective of the analysis is to develop the set of conditions for correct and successful motion segmentation for each type of motion using the fundamental matrix motion model.

The main contribution of this thesis is the theoretical analysis of the feasibility of motion segmentation in commonly encountered scenarios including:

1. motion-background segmentation,
2. segmentation of translational motion, and
3. segmentation of planar motion

in dynamic scenes using the fundamental matrix motion model.

1.3 Overview of the thesis

The thesis is divided into seven chapters including the introduction. The rest of the thesis is organised as follows:

Chapter 2 starts by describing the building blocks of a motion-segmentation algorithm and provides a critical review of previous works in motion segmentation using the fundamental matrix. It introduces the concept of image data, camera model and feature extraction. Furthermore, the concept of a fundamental matrix motion model and its approximations for special motions, i.e. static object (zero motion), translational and planar motions, are explained. The discussion continues with a review of existing estimation methods and error measures used to estimate a fundamental matrix. Finally, the segmentation strategies and robust estimation techniques used in motion segmentation are described.

Chapter 3 defines the scope and methodology of the feasibility analysis of motion segmentation using the fundamental matrix. The analysis is divided into three main parts based on the types of motions and approximations of the fundamental matrix. Each part consists of dynamic scene modelling and the theoretical derivation of the conditions for successful motion segmentation. The derived conditions are verified via experiments using both synthetic and real-image data.

Chapter 4, **5** and **6** detail the theoretical derivation for successful motion-background segmentation, translational-motion and planar-motion segmentations, respectively. The conditions for segmentation are examined when all motions and scene parameters are varied. The derived conditions are verified via experiments, the design of which is based on the Monte Carlo method, using synthetic images. Experiments

using real-image data are applied to demonstrate the capability of the proposed conditions to correctly predict the outcome of motion segmentation.

Chapter 7 concludes the analysis of motion segmentations and reviews the contributions of the research work. In addition, several recommended future works and research directions are discussed.

Chapter 2

Motion Segmentation Using Fundamental Matrix: A Review

Motion segmentation involving multibody structure-and-motion (MSaM) is a complex process involving three main tasks [123], which are to extract all feature points from 2D images, to estimate the motion model and to decide on the *inlier-outlier* dichotomy. In practice, the unknown camera and motion parameters in most applications require the usage of the most general motion model which is the fundamental matrix. In addition, a motion-segmentation process needs to be robust to tolerate potential wrong matches from feature extraction and contamination of noise in the data.

This chapter reviews several related works in motion segmentation using the fundamental matrix (section 2.1) and the theoretical background for all stages of the segmentation process — namely, feature extraction (section 2.2), application of the fundamental matrix motion model (section 2.3) and segmentation strategy (section

2.4) — to understand the theories behind them and their practical limitations.

2.1 Related works

The solutions for motion segmentation using the fundamental matrix and its approximates can be broadly classified into non-algebraic methods (Shapiro et al. [82, 83], Torr et al. [94, 104] and Schindler et al. [79]), and algebraic methods (Wolf and Shashua [120], Vidal et al. [107, 109, 110, 111, 113], and Vidal and Ma [108]). Algebraic methods use the algebraic constraint satisfied by all objects in the scene whereas in non-algebraic methods, detection and recovering of each SaM is usually an iterative process [81].

One of the first modern techniques for the recovery of 3D motions was developed by Shapiro et al. [82, 83], in which all image points were considered simultaneously and the motions were modelled using the affine approximation of the fundamental matrix. The performance of this approach in terms of its accuracy, tolerance to noise and matching errors or *gross outliers* were better then the segmentation approaches that required data sampling [82, 83].

Then, Torr et al. [94, 102, 104] presented a way to automatically determine the number of motions and the appropriate motion model for each motion in a particular scene. In their work, the use of seven motion models was proposed to eliminate the possibility of non-unique solutions while estimating the motion model due to the *degeneracy* problem [95, 104]. These models included the fundamental matrix, the affine and the translational fundamental matrix and also the plane homography. After the number of motions was determined, the parameters for each motion were

estimated by alternating between feature-clustering and motion estimation in a probabilistic framework using the Expectation Maximisation (EM) algorithm [94]. The main issue with the proposed algorithm was the dependence of the EM algorithm on its initialisation procedure [102].

Wolf and Shashua [120] presented a method for two-body motion segmentation that uses a geometrical constraint derived from all image points and called it the *segmentation matrix*. Using this matrix, they could estimate the fundamental matrix, its affine approximation or the translational fundamental matrix associated with each body [120]. This method has the advantage of being able to determine the camera parameters under affine camera assumptions by recovering the homography at infinity when the relative motion between the two bodies is a pure translation [120].

Vidal et al. developed an algebraic method called the Generalized Principal Component Analysis (GPCA) for solving the problem of data-fitting to a linear subspace with an unknown dimension [107, 109, 110]. The GPCA starts by algebraically determining the number of unique subspaces to represent the data and their dimensions, and then decides on the segmentation of the data. Since the GPCA estimates the dimension of each subspace in the data, it is applicable to a wide range of subspaces or motion models, including the fundamental matrix with respect to motion-segmentation problems. The advantage of this approach is that it algebraically solves the motion model and is thus able to eliminate the feature-clustering stage in motion-segmentation problems [112]. The main issue with this approach is the huge amount of data points required by the GPCA when dealing with multiple subspaces with high dimensions.

As a specific solution to motion-estimation and segmentation problems, Vidal et al. [111] derived and proposed the multibody fundamental matrix; the generalisation of the epipolar constraints or the fundamental matrix for multiple motions. This work was then extended to also work with a two-dimensional motion model based on the optical flow [108]. Recently, Schindler and Suter have used the multibody fundamental matrix for two-view MSaM segmentation and have improved the segmentation performance by implementing the geometric model selection to replace degenerate motion/s in dynamic scenes [80, 81].

Klappstein, Stein and Franke studied the detection of moving objects for the applications of local navigation and automobile driver-assisted systems [51]. Several motion models over two and three image frames were considered, including the fundamental matrix, positive depth and height constraint and the trifocal tensor. However, the detection analyses were limited to circular, parallel and lateral motions, which are usually seen from a monocular camera mounted in front of an automobile.

Even though, the analyses aimed at solving motion-segmentation problems have received considerable attention over the years (summarised in [25, 39, 61]), the conditions governing the feasibility of detection and segmentation of each structure-and-motion in a dynamic scene are largely unaddressed. These conditions are able to provide practical limitations of motion-segmentation solutions and serve as useful guidelines for practitioners designing motion-segmentation solutions for computer-vision applications.

2.2 Image formation and feature extraction

An image is a map of a 3D scene on a 2D plane. In terms of computer vision, it is a transformation from 3D scenes to 2D images. The fundamental concept of the 3D-to-2D transformation is based on the theory of perspective projection using a pinhole camera model as shown in figure 2.1. The relationship between the homogenous

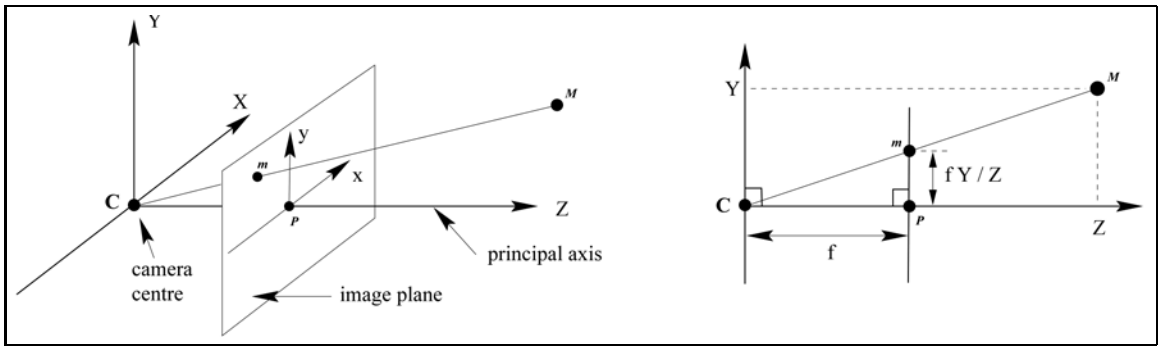


Figure 2.1: The pinhole camera model. Figure is from [39].

coordinate of a 3D point $M = [X\ Y\ Z\ 1]^\top$ in figure 2.1 and its equivalent point on a 2D image $m = [x\ y\ 1]^\top$ can be expressed as:

$$m = A[R \mid T]M, \tag{2.1}$$

where matrices A and $[R \mid T]$ summarise the intrinsic and extrinsic parameters of the camera [23, 39]. The 3×3 matrix A is called the *camera calibration matrix*

$$A = \begin{bmatrix} f & 0 & P_x \\ 0 & f & P_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{2.2}$$

with the symbol f denoting the camera focal length and the symbol $[P_x\ P_y]$ denoting the location of the principal point of the image. The matrix $[R \mid T]$ in equation

(2.1) represents changes in camera position by rotation R followed by a translation T. For the case of a static camera, the matrix $[R \mid T]$ reduces to $[I \mid 0]$, which is a combination of an identity matrix and a vector of zeros.

Besides the perspective projection and pinhole camera model, shown in figure 2.1, there are many other camera models that have been developed, including the affine camera model [70] and the pushbroom camera model [31]. For a review of these models and their theoretical derivations, the reader is referred to [2, 23, 39].

The image data contain rich and high-density information about a particular scene such as colour, texture, object size and position, and object motion in the case of multiple images. In many computer-vision applications, this information is redundant; thus, for optimisation purposes, a feature detector is used to filter out some of the irrelevant information. In terms of motion estimation and segmentation, a feature detector is applied to extract point correspondences from two or more images to represent object motions.

Generally, a feature detector functions by first determining the data primitives, such as individual pixels, corners, lines, blocks, blobs and T-junctions [88]. Then these features are represented by unique feature vectors and finally are matched with their correspondences from different images. The matching is done by solving a particular cost function in terms of distances between those vectors. Importantly, a feature detector for motion-segmentation problems needs to be repeatable and reliable under different viewing conditions, and also scale-invariant to handle changes in the size of objects in the image.

One of the most popular feature detectors developed for computer-vision problems is the Harris corner detector [32]. However, the Harris corner detector can not handle

variations of image size or image-scaling which are critical in motion-segmentation applications. The image-scaling problems are solved by the automatic scale selection algorithm developed by Lindeberg [54] and later improved by Mikolajczyk and Schmid [66] to produce a highly repeatable and robust feature detector. Several other feature detectors for computer-vision problems are also proposed — namely, the Scale-Invariant Feature Transform (SIFT) algorithm [57, 58] and the Speeded-Up Robust Features (SURF) algorithm [11] — mainly to improve their repeatability, robustness, stability and computational speed [49, 50]. For a review of these feature detectors and a comparison of their performance, the reader is referred to [67, 68, 72].

In this work, we consider a case where a static and uncalibrated camera has a camera matrix according to equation (2.1) and a feature detector based on the SIFT algorithm [58]. This is because they are widely used in many computer-vision applications and the SIFT algorithm has been shown to outperform other feature detectors in terms of repeatability, robustness and ability to tolerate image scale variation [68].

2.3 The fundamental matrix motion model

Over the years, the computer-vision community has developed a number of motion models to represent moving objects in a dynamic scene [25, 39, 61]. Common examples are: the fundamental matrix for a rigid 3D object having an arbitrary motion; plane homography [24, 84] for moving planar object; and 2D optical flow [1, 15, 89] or models based on change-detection [63, 99] to represent apparent motions on the image planes.

The analysis in this work focuses on motion segmentation using the most gen-

eral motion model for 3D SaM, i.e. the fundamental matrix. The theory of the fundamental matrix motion model, introduced in [26, 27, 37, 33], is an improvement on the novel essential matrix [55] for cases involving an uncalibrated camera. The fundamental matrix summarises the intrinsic projective geometry between two views [39] and only depends on the camera parameters and its position [39]. In terms of motion-segmentation problems, the fundamental matrix takes into account all scene and motion parameters, i.e camera parameters and 3D motion, size and location of object [123]. In practice, the fundamental matrix motion model represents a more realistic motion model compared to the models based on apparent changes on the image planes, i.e. the 2D optical flow or change detection [123].

In addition, the fundamental matrix is suitable for many computer-vision applications since it does not require prior knowledge of the exact nature of object motions and camera parameters [14, 104]. This results in the fundamental matrix being one of the preferred motion models in motion-segmentation applications because it will not be adversely affected by small changes to the camera parameters that occurr due to lens focusing or camera vibrations [14].

Consider a scene in figure 2.2 where an uncalibrated camera is moved according to rotation R followed by translation T from position C_1 to position C_2 while taking two images of a point $M_i = [X_i\ Y_i\ Z_i\ 1]^\top$ in 3D-space from each position. The corresponding image/feature points associated with a point M_i for both camera positions are given by the homogenous coordinates $m_{1i} = [x_{1i}\ y_{1i}\ 1]^\top$ and $m_{2i} = [x_{2i}\ y_{2i}\ 1]^\top$, respectively. The relationship between the points m_{1i} and m_{2i} on two images in figure

2.2 are according to equation

$$\begin{bmatrix} x_{2i} & y_{2i} & 1 \end{bmatrix} F \begin{bmatrix} x_{1i} \\ y_{1i} \\ 1 \end{bmatrix} = 0, \tag{2.3}$$

where

$$F = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ h_7 & h_8 & h_9 \end{bmatrix}, \tag{2.4}$$

is a unique 3×3 rank-2 matrix called the *fundamental matrix* [25, 39, 61]. The fundamental matrix can be computed using equation

$$F = A^{-T}[T]_x R A^{-1}, \tag{2.5}$$

where $[T]_x$ is a *skew symmetric* matrix of the translation T [3, 39, 124].

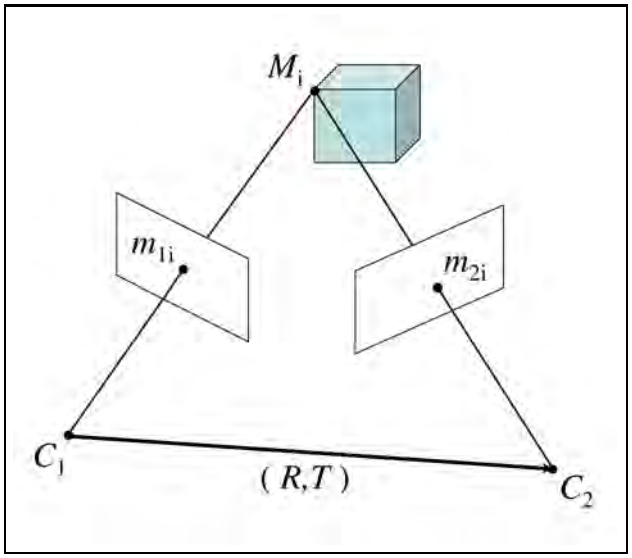


Figure 2.2: An uncalibrated scene for the fundamental matrix motion model. Figure is from [61].

Meanwhile, if the camera matrix A is known, i.e. in the calibrated case, the equation (2.3) is reduced to:

$$\begin{bmatrix} x_{2i} & y_{2i} & 1 \end{bmatrix} E \begin{bmatrix} x_{1i} \\ y_{1i} \\ 1 \end{bmatrix} = 0, \tag{2.6}$$

where E is a 3×3 essential matrix based on the external camera parameters or, specifically, the motion of the camera [55]:

$$E = [T]_x R. \tag{2.7}$$

The role of the moving camera and static object in figure 2.2 could be interchanged based on the principle of duality [17, 117]. Thus, the fundamental matrix in equation (2.5) is also applicable for the case of motion segmentation where a static camera is observing objects in motion [61]. For the complete derivation of the essential and the fundamental matrices using both geometric and algebraic methods, the reader is referred to [39, 55, 61, 121].

2.3.1 Fundamental matrix estimation

In a practical motion-segmentation problem, the fundamental matrix associated with a particular motion needs to be estimated from corresponding image/feature points. This is due to the fact that the camera parameters and object motion are not known, *a priori*. The basic equation for the estimation of a fundamental matrix is derived by expanding equation (2.3):

$$x_{2i}x_{1i}h_1 + x_{2i}y_{1i}h_2 + x_{2i}h_3 + y_{2i}x_{1i}h_4 + y_{2i}y_{1i}h_5 + y_{2i}h_6 + x_{1i}h_7 + y_{1i}h_8 + h_9 = 0, \tag{2.8}$$

which is factorised to:

$$(x_{2i}x_{1i} + x_{2i}y_{1i} + x_{2i} + y_{2i}x_{1i} + y_{2i}y_{1i} + y_{2i} + x_{1i} + y_{1i} + 1)\mathbf{h} = 0, \tag{2.9}$$

where $\mathbf{h}$ is the 9×1 vector made up of the elements of F in (2.4).

Several methods have been developed to estimate the fundamental matrix from image-point correspondences and these are classified as linear, iterative and robust methods [3, 124]. The essence of the linear methods is that they estimate the fundamental matrix by solving equation 2.9 using either seven or eight corresponding image points [3, 124]. The fundamental matrix estimate can be obtained by solving equation (2.9) algebraically using at least seven corresponding image points via singular-value decomposition (SVD) and the rank-2 constraint of the fundamental matrix [29, 35, 96].

In a linear method, called the *eight-point algorithm* [36, 55], the fundamental matrix is estimated by minimising the sum of squares of algebraic distance:

$$\min_{F} \sum_{i} (m_{2i}^{\top} F m_{1i})^2, \tag{2.10}$$

using either least-squares [60] or orthogonal least-squares techniques [101]. The linear methods are computationally efficient, however their accuracies are poor, especially in the presence of wrong matches from feature extraction and bad locations of image points due to contamination from measurement noise [3].

The iterative methods for the estimation of a fundamental matrix involve minimising a cost function in terms of distances between corresponding image points to an eight-dimensional manifold representing a model candidate for the fundamental matrix estimate. The iteration is repeated until the cost function is minimised, which

means that the best candidate for the fundamental matrix estimate is found. Iterative methods are capable of producing superior accuracies of fundamental matrix estimates compared to linear methods; however, they are still unable to handle wrong matches [3, 124].

Three commonly used cost functions for the estimation of a fundamental matrix using iterative methods are:

1. The sum of squares of geometric distances [39]

$$\min_F \sum_i \left[(x_{1i} - \hat{x}_{1i})^2 + (x_{2i} - \hat{x}_{2i})^2 + (y_{1i} - \hat{y}_{1i})^2 + (y_{2i} - \hat{y}_{2i})^2\right]^2, \qquad (2.11)$$

where $\hat{x}_{1i}$, $\hat{x}_{2i}$, $\hat{y}_{1i}$ and $\hat{y}_{2i}$ are the estimated locations of image points $\hat{m}_{1i}$ and $\hat{m}_{2i}$ that satisfy $\hat{m}_{2i}^\top \hat{F} \hat{m}_{1i} = 0$ calculated using the candidate of the fundamental matrix estimate $\hat{F}$. This measure minimises the distances between corresponding image points to their reprojection using the $\hat{F}$.

2. The sum of squares of Sampson distances [101, 119]

$$\min_F \sum_i \left[\frac{m_{2i}^\top F m_{1i}}{\sqrt{[(\frac{\partial}{\partial x_{1i}})^2 + (\frac{\partial}{\partial y_{1i}})^2 + (\frac{\partial}{\partial x_{2i}})^2 + (\frac{\partial}{\partial y_{2i}})^2] m_{2i}^\top F m_{1i}}}\right]^2. \qquad (2.12)$$

The Sampson-distance measure is the first-order approximation of the geometric distance in equation (2.11) with high accuracy, i.e. up to 4 or 5 significant figures [101].

3. The sum of squares of Luong distances [59]

$$\min_F \sum_i \left[\frac{m_{2i}^\top F m_{1i}}{\sqrt{[(\frac{\partial}{\partial x_{1i}})^2 + (\frac{\partial}{\partial y_{1i}})^2] m_{2i}^\top F m_{1i}}} + \frac{m_{2i}^\top F m_{1i}}{\sqrt{[(\frac{\partial}{\partial x_{2i}})^2 + (\frac{\partial}{\partial y_{2i}})^2] m_{2i}^\top F m_{1i}}}\right]^2. \qquad (2.13)$$

The Luong-distance measure is also an approximation of the geometric distance. It differs from the Sampson distance by computing the distances separately in each image. However, the use of Luong distance produces slightly inferior estimates compared to the Sampson-distance measure [39, 124].

In this analysis, we use the cost function based on the Sampson-distance measure [119, 101] due to its practicality in terms of lower computing requirement and high accuracy compared to the geometric-distance [104, 101] and the Luong-distance measure [39, 124].

The robust methods for the estimation of the fundamental matrix incorporate robust statistical-regression techniques to handle both wrong matches and contamination of measurement noise. These methods are usually preferred since they represent the real problem encountered in many computer-vision applications involving fundamental matrix estimation[1]. For a complete review of all methods for the estimation of a fundamental matrix and their performances, the reader is referred to [3, 101, 124].

2.3.2 Degeneracy and fundamental matrices for special motions

The degeneracy problem occurs when a set of corresponding image points produce non-unique estimates of the fundamental matrix [39, 104]. In this case multiple solutions of the fundamental matrix can be estimated from those points. This condition is not desired as it could produce errors in the estimation of the motion model which results in an incorrect *inlier-outlier* dichotomy with respect to motion-segmentation

[1]More details about robust estimators are covered in section 2.4.

problems.

The problem of degeneracy in the fundamental matrix is caused by either structure degeneracy or motion degeneracy [98]. Structure degeneracy takes place when the corresponding points are coplanar or coming from the same plane. This situation is usually encountered when the image points associated with a 3D object have a relatively small depth compared to the distance between the object and camera, for example, a camera viewing a small building located very far away where most of the image points appear to be on the same plane. In such cases, the image data does not contain enough information in terms of the depth for the fundamental matrix motion model and a plane homography is a much more appropriate model [62].

The problem of motion degeneracy occurs when the camera or object has a restricted motion which is less than six degrees-of-freedom since the fundamental matrix takes into account all motion parameters, i.e. rotation and translation around or along X, Y and Z axes in a 3D-space. Common examples occur in the situations where a camera does not translate or only rotates around its center. In these situations, the relationship between the image points in the two images would also be best described by a plane homography, as the fundamental matrix is a zero matrix [98], according to equation (2.5) when the translation T is zero.

In addition, there are two other fundamental matrix approximations to represent restricted motions and also to eliminate the possibility of degeneracy. There are: the translational fundamental matrix for camera or object having a pure translational

motion without any rotation

$$F_T = \begin{bmatrix} 0 & q_3 & -q_2 \\ -q_3 & 0 & q_1 \\ q_2 & -q_1 & 0 \end{bmatrix}, \tag{2.14}$$

and the affine fundamental matrix [13, 114]

$$F_A = \begin{bmatrix} 0 & 0 & r_1 \\ 0 & 0 & r_2 \\ r_3 & r_4 & r_5 \end{bmatrix}, \tag{2.15}$$

which arises from a scene viewed by an affine camera [39, 70] or a 3D object having planar motion [104], i.e. a motion restricted to a plane orthogonal to the camera optical axis.

These fundamental matrix approximations are important because they are commonly encountered in many practical computer-vision applications [39]. For example, a zero fundamental matrix arises from images of a static object in the background, or an object having a pure rotation, such as in a turntable sequence. Pure translational motions are encountered in many traffic surveillance applications while planar motions are seen in a scene when the ratio of the object motions along the camera optical axis over the distance between the objects and the camera is negligible [70, 83, 101, 121].

In practice, to ensure that the estimated motion model does not degenerate, a model-selection algorithm is usually implemented to select the most suitable motion model to represent a particular motion in a dynamic scene [80, 81]. Model selection problem is however outside the scope of this thesis and is not explained here.

2.4 Segmentation strategies and robust estimation

A robust estimator is a regression technique based on statistical theory and is used to estimate certain parameters from a particular data set. In terms of motion segmentation, a robust estimator is commonly used to estimate the motion model and to decide on the *inlier-outlier* dichotomy of the data. This is because many motion-segmentation applications are dealing with data set containing multiple structures and motions, wrong matches or *gross outliers* and errors from the contamination of measurement noise. In general, robust estimators used to solve computer-vision problems include three main steps [43]:

1. Optimisation: return an initial estimate of the motion model, i.e. a fundamental matrix for a 3D SaM, as a result of the optimization of a cost function. For the fundamental matrix motion model, the cost functions are in terms of distances provided in equations (2.10) to (2.13).

2. Segmentation: decide on the separation between *inlier* and *outlier* by evaluating their magnitude of distance to the manifold given by an estimate of the motion model. The scale estimate of the *inliers* population determines the threshold for deciding the *inlier-outlier* dichotomy. In other words, all points with their associated distances larger than the threshold are considered as the *outliers*.

3. Refinement: refine the estimate of the motion model by applying the least-squares technique only to the *inlier* population detected in the segmentation step.

Many robust estimators use random sampling in their optimisation step in order to find the best candidates in the data by satisfying a particular cost function. In terms of fundamental matrix estimation, a robust estimator searches for seven or eight image/feature points based on equation (2.9) that optimise a particular distance measure (equations (2.10) to (2.13)). The probability of a robust estimator finding at least all candidates associated with the target motion/structure or *inlier* using random sampling is given by the equation:

$$P_{\text{success}} = 1 - [1 - \epsilon^p]^B. \tag{2.16}$$

The symbol ϵ is the ratio of *inliers* in the data set, p is the dimension of the model and B is the number of trials. Thus, to ensure an accurate estimation of the motion model with the probability of P_{success}, the random sampling needs to be repeated for B iterations:

$$B = \frac{\log(1 - P_{\text{success}})}{\log(1 - \epsilon^p)}. \tag{2.17}$$

This shows that the random sampling process requires a high computational load, making it undesirable for real-time applications. The number of samples B in equation (2.17) increases exponentially when the motion model has high dimensions, i.e. seven or eight dimensions for the fundamental matrix motion model, and the smallest possible value of inlier ratio ϵ is usually used since it is not known in advance for most applications.

To overcome the high computational cost associated with random sampling, guided sampling techniques were applied to reduce the number of samples when the information about the reliability of the data set was roughly known. This information could come either from user input or be heuristically estimated from the data itself

[18, 43]. The essence of a guided sampling technique is that it minimises the number of samples by selecting data candidates which have a high probability of being the *inlier* to the data [93, 97]. There are a number of techniques for guided sampling which have been developed for robust estimators such as in [4, 21, 30, 43, 45, 71, 93]; however, in practice, the information about the reliability of the data is not usually known in advance for most motion-segmentation applications.

The performance of a robust estimator is usually measured by its *breakdown point* which is defined as the minimum percentage of *outliers* contamination that can cause an estimator to produce arbitrarily large values [77]. For example, if the *breakdown point* of a particular estimator is around 40%, it means that the estimator should be able to produce a correct parameter estimate if *outliers* contamination is less than 40% of the entire data. However, the performance of a robust estimator does not necessarily live up to its expected *breakdown point* if the size of the data set is relatively small [44]; this usually occurs in a scene containing objects with a small number of detectable features. This is because the accuracy of the scale estimate for the *inliers* deteriorates when the size of the *inliers* population in the data set is small [44].

Robust estimators have a very long history and have been used in computer vision for more that a quarter of a decade. In the early stage, estimators are mainly developed for statistical applications assuming that only a single structure or model exists in the data. A common example is the Least-Square (LS) technique — based on the theory of linear regression — which performs the parameters estimation by minimising the sum of squared residuals [77]. The main issues with the LS technique are its inability to handle the *outliers* in the data and its *breakdown point* of 0%.

In order to handle the presence of *outliers* in the data, several estimators have been designed, namely, the least-median-of-squares (LMedS) [76] and the family of maximum likelihood, or M-estimators [41, 47, 48, 86]. The LMedS minimises the median of the residuals and theoretically it has a *breakdown point* of 50% [76] whereas, the M-estimator uses a symmetric function to reduce the effect of *outliers* in the data [47]. Even though LMedS was mainly developed for statistical applications, it has been used in a number of computer-vision applications due to its robustness and relatively high *breakdown point*. Examples of applications of LMedS in computer vision are in solving pose-estimation problems [20, 74], optical-flow calculations [5, 73], range-image segmentation [75] and object recognition and tracking [106]. The main issue with LMedS is its performance deterioration when the data is contaminated by Gaussian noise [77].

In practice, a robust estimator for computer-vision applications requires a *breakdown point* of larger than 50% since it needs to be able to process a data set with a majority of *outliers*; especially in cases involving multiple structures or objects in a scene [64]. To meet this requirement, a number of robust estimators have been specially designed for computer-vision problems where the *breakdown point* is much larger than 50%. For example, the Hough transform [46], the Random Sample Consensus (RANSAC) [28], the Residual Consensus (RESC) [122], the projection-based M-estimators [18, 90, 91], the Minimise the Probability of Randomness (MINPRAN) [85], the Minimum Unbiased Scale Estimator (MUSE) [69], the Maximum Likelihood Estimation Sample Consensus (MLESAC) [103], the Adaptive Least k^{th} Order Squares (ALKS) [53], the Modified Selective Statistical Estimator (MSSE) [6], the High Breakdown M-estimator (HBM) [43], the Two-Step Scale Estimator (TSSE)

and the Adaptive Scale Sample Consensus (ASSC) [115, 116].

Most of the above estimators have been successfully applied to computer-vision problems; to name a few, motion segmentation [100, 105] and structure-from-motion [12, 34, 92, 125]. For a theoretical analysis of robust estimation for data including two distinct populations and a complete review of robust estimators, the reader is referred to [19, 64, 65, 86, 87].

In this work, the segmentation step is based up on using MSSE [6]. The choice is motivated by the MSSE desired performance in terms of consistency [42] and the fact that it has been shown to be successful in segmenting closely-spaced structures [40, 44]. It is important to note that, although the segmentation steps within MSSE are used, all of the analysis is general and similar results would be obtained if other robust estimators are used.

2.5 Conclusion

In this chapter, we have critically reviewed the theories and methods for motion segmentation using the fundamental matrix, including their practical limitations. Although the estimation of a fundamental matrix and its use for motion segmentation are well understood, the conditions governing the feasibility of detection and segmentation of each structure-and-motion are largely unaddressed. These conditions are important as they provide useful guidelines for practitioners designing motion-segmentation or estimation solutions for computer-vision problems. In the next chapter, we will provide the details of the methodology adopted to analyse the feasibility of motion segmentation using the fundamental matrix for different types of motions.

Chapter 3

Scope and Methodology

This chapter presents the scope and methodology for a feasibility analysis of motion segmentation. The analysis starts with modelling a dynamic scene, presented in section 3.1, including two 3D-rigid objects having distinct motions. In section 3.2, motion segmentation is performed to determine each object in the scene and to derive the theoretical limits for successful segmentation. The focus is on developing measures for the degree of separation between motions using the fundamental matrix motion model. Based on these measures, a set of conditions to guarantee successful motion segmentation is proposed. Section 3.3 describes the experiments, using synthetic images, which are designed to evaluate the validity of the proposed conditions and examine their effect on variations of scene and motion parameters. The applicability of the proposed conditions and their relevance to the problems encountered in real motion-segmentation applications are demonstrated in experiments using real-image data, described in section 3.4. Finally, section 3.5 concludes this chapter.

The feasibility analysis is organised based on the types of motions in a dynamic

scene and the approximate models of the fundamental matrix to eliminate the possibility of degeneracy, when dealing with less general motion, i.e. motion less than six degree-of-freedom [25, 39]. In practice, the less general motions are usually seen in a number of computer-vision applications. For example, static points are associated with the background or objects very far from the camera, pure translational motions are present in most traffic surveillance applications and planar motions are present in a scene where the motions along the camera optical axis are small compared to the object-to-camera distance [39, 70, 83, 101, 121]. Therefore we have divided the analysis into three main parts:

1. motion-background segmentation in chapter 4,

2. translational-motion segmentation using the translational fundamental matrix in chapter 5, and

3. planar-motion segmentation using an affine fundamental matrix in chapter 6.

3.1 Modelling a dynamic scene

To analyse the motion-segmentation problem, we consider the general case of a stationary and uncalibrated camera viewing a dynamic scene including two rigid 3D-objects that move according to two distinct motions denoted by motion-a and motion-b. Both motion-a and motion-b are parameterised by rotation θ and followed by a translation T i.e. θ_a and T_a for motion-a and θ_b and T_b for motion-b where $T_a = [T_{xa}\ T_{ya}\ T_{za}]^\top$ and $T_b = [T_{xb}\ T_{yb}\ T_{zb}]^\top$.

The focus on only two motions present in a scene is justified because the segmentation is an iterative process and the analysis aims to find the smallest amount of relative motion that can be detectable. Intuitively, in a motion-segmentation problem involving more than two motions, the motion that includes the highest number of points will be segmented first and then the segmentation process is repeated for other motions in the scene. In addition, the assumption of a stationary camera viewing the scene is justified, since only the motions relative to the camera is relevant to this analysis [100].

Consider a point in the 3D-space with coordinates $M_i = [X_i \; Y_i \; Z_i]^\top$ viewed by a camera with a camera matrix A

$$A = \begin{bmatrix} f & 0 & P_x \\ 0 & f & P_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3.1}$$

with focal length f and the image principal point $[P_x \; P_y]$, and denote its corresponding point on the image plane by $\underline{m}_{1i} = [\underline{x}_{1i} \; \underline{y}_{1i}]^\top$. After an arbitrary motion, image point $\underline{m}_{1i}$ moves to position $\underline{m}_{2i} = [\underline{x}_{2i} \; \underline{y}_{2i}]^\top$ on another image.

The images contain $N_i + N_o$ points where N_i and N_o denote the number of points having either motion-a and motion-b, respectively. The number of image points can be controlled by varying the *inlier* ratio ϵ:

$$\epsilon = \frac{N_i}{N_i + N_o}. \tag{3.2}$$

We assumed that there were no mismatches or *gross outliers* in the images to eliminate their effect on the analysis and segmentation results. In practice, *gross outliers* are removed by the robust estimator commonly used as part of the motion-segmentation

process [123].

All points on the image planes are contaminated by measurement noise assumed to be independently and identically distributed (i.i.d) with Gaussian distribution:

$$\begin{aligned} x_{1i} &= \underline{x}_{1i} + e^1_{ix}, & y_{1i} &= \underline{y}_{1i} + e^1_{iy}, \\ x_{2i} &= \underline{x}_{2i} + e^2_{ix}, \quad \text{and} & y_{2i} &= \underline{y}_{2i} + e^2_{iy}, \end{aligned} \tag{3.3}$$

where $[e^1_{ix}\ e^1_{iy}\ e^2_{ix}\ e^2_{iy}]^\top \sim N(0, \sigma_n^2 I_4)$ in which σ_n is the unknown scale of noise and I_4 is a 4×4 identity matrix. The underlined variables denote the true noise-free locations of the points in the image planes.

Without loss of generality, motion-a is considered as the target motion and motion-b as the unwanted one. In the context of robust estimation, the matching points associated with motion-a are assumed to be *inliers*, which we aim to segment from the matching points having motion-b, i.e. *outliers*.

3.2 Motion segmentation using fundamental matrix

The analysis of motion segmentation aims to derive a measure for the degree of separation between two motions — motion-a and motion-b where motion-a is the target motion (*inliers*) — using fundamental matrix motion model. The derived measure is used as the basis to determine a set of conditions to guarantee successful segmentation.

The analysis focuses on investigating the theoretical limit of motion segmentation, in terms of obtainable scene and motion parameters, and how imperfect estimate of

the fundamental matrix would affect the conditions for segmentation is beyond the scope of this work. In practice, the fundamental matrix can be accurately estimated using a number of robust methods proposed in computer-vision literatures [3, 101, 124] and the estimation issues in terms of both the feasibility and the accuracy, has already been thoroughly studied [42, 44]. Thus, in our analysis we assume that an accurate estimate of the fundamental matrix associated with the target motion is provided by a robust estimator. As such the fundamental matrix of motion-a is calculated using equation (2.5).

To decide on the *inlier-outlier* dichotomy in the mixture of all image points, the segmentation steps within the Modified Selective Statistical Estimator (MSSE) [6] is used due to its desired performance in terms of consistency [42] and that it has been shown to be successful in segmenting closely spaced structures [40, 44]. It is important to note that, although we use MSSE in our experiments, the analysis is general and similar results would be obtained if other robust estimators are used.

The residual for motion segmentation is the square of the Sampson-distance measure [101, 119]:

$$d_i = \frac{m_{2i}^{\top} F m_{1i}}{\sqrt{[(\frac{\partial}{\partial x_{1i}})^2 + (\frac{\partial}{\partial y_{1i}})^2 + (\frac{\partial}{\partial x_{2i}})^2 + (\frac{\partial}{\partial y_{2i}})^2] m_{2i}^{\top} F m_{1i}}}, \tag{3.4}$$

and d_i denote the distance of the i^{th} point from the motion manifold with respect to the fundamental matrix of motion-a. The Sampson-distance measure is applied because it is commonly used due to its lower computing requirement and high accuracy (up to 4 or 5 significant figures [101]) compared to the geometric distance measure [101, 104]. The Sampson-distance measure also produces slightly better results than the Luong-distance measure in equation (2.13) [39, 124].

In MSSE, the residuals (${d_i}^2$) are sorted in an ascending order, indexed by the symbol k, and an unbiased estimate for the scale of noise, given by k points with the smallest distances, is [6]:

$$\sigma_k^2 = \frac{\sum_{i=1}^{k} d_i^2}{k-1}. \tag{3.5}$$

While incrementing the value of k, the *inlier-outlier* dichotomy occurs as soon as the magnitude of d_{k+1} becomes larger than 2.5 times the scale estimate given by the smallest k distances:

$$|d_{k+1}| > 2.5\sigma_k. \tag{3.6}$$

With the threshold of 2.5, at least 99.4% of points associated with motion-a (*inliers*) will be correctly segmented if their d_i are normally distributed [6]. In practice, the measurement values are always bounded and the above threshold would represent a perfect segmentation.

Equations (3.5) and (3.6) show that the success of motion segmentation depends on the values of the residuals or Sampson distances d_i associated with both motions. If the residuals of points having motion-a are sufficiently different from the residuals associated with points having motion-b, the segmentation is expected to be successful. A measure for the degree of separation between two motions based on the relative distance between both populations of d_i will be introduced in this thesis. The condition for motion segmentation is determined based on the degree of separation to ensure that the distribution of residuals are always sufficiently far from each other. If the condition for segmentation is satisfied, the segmentation is guaranteed to be successful. In practice, the outcome of a motion segmentation could be predicted via the value of the degree of separation between two motions estimated from obtainable scene and

motion parameters. Therefore, this condition serves as a guideline for practitioners in designing motion segmentation solutions for computer-vision applications.

3.3 Monte Carlo experiments using synthetic images

The theoretical analysis in section 3.2 is verified via experiments using synthetic images. The foci are:

1. to examine the validity of the proposed degree of separation between two motions and

2. to determine the conditions for segmentation when all scene and motion parameters are varied, i.e. *inlier* ratio, camera parameters, objects motions, sizes and locations.

The experiments are designed to represent identical conditions in the theoretical analysis and are based on the Monte Carlo statistical method.

In the experiments, 1000 segmentations are performed for each scene parameter (*inlier* ratio, camera parameters, objects sizes and locations) while the motion parameters are randomly selected based on the values of the degree of separation between two motions. The segmentation performance is measured by the ratio (ζ) of the number of segmented points having motion-a over the true number of points having motion-a. The value of ζ equal to one signifies correct segmentation while ζ larger than one means over-segmentation, which indicates that some points having motion-b are also segmented as motion-a due to the similarity between the two motions. To

examine the consistency of the segmentation results, the statistical mean and standard deviation of 1000 ζs, denoted by $\bar{\zeta}$ and σ_ζ, are also calculated. We considered the segmentation of motion-a to be correct and consistent if $\bar{\zeta} \approx 1$ and $\sigma_\zeta \leq 0.01$ throughout the experiments.

In addition, using the above experiments, the dominant scene or motion-parameters affecting the performance of a motion segmentation can be identified and examined. The experimental results are presented in terms of the conditions for motion-background segmentation, translational-motion and planar-motion segmentation — the proposed conditions are based on the derived measures for the degree of separation between two motions.

3.4 Experiments using real-image data

Experiments using real-image data are designed to show the relevance of the analysis and Monte Carlo experiments in section 3.2 and 3.3 to the problem encountered in a real motion-segmentation problem. In addition, the experiments aim to demonstrate the capability of the proposed conditions to correctly predict the outcome of a set of segmentation scenarios.

In the experiments, motion segmentation is performed in cases where the values of the degree of separation between two motions are less than, close to and larger than the proposed thresholds for successful segmentation. The segmentation performance and the histogram of distances associated with all points for each case are examined and compared with the results from experiments using synthetic images.

3.5 Conclusion

The feasibility analysis of motion segmentations is organised in the following steps:

1. development of a model to represent a dynamic scene,

2. theoretical derivation of quantitative measures for the degree of separation in motion-background segmentation, translational-motion segmentation and planar-motion segmentation,

3. Monte Carlo experiments to verify the theoretical analysis and develop the conditions for successful segmentation, based on the derived measures for the degree of separation between two motions, and

4. experiments using real-image data to show the relevance between the analysis and the problems encountered in real-image applications.

In practice, the value of the degree of separation between two motions could be estimated using obtainable scene and motion parameters. By comparing the estimated value of the degree of separation with the proposed condition for segmentation, the outcome of motion segmentation could be predicted. Thus, these conditions serve as a guideline for practitioners designing motion-segmentation solutions for computer-vision problems. In the following chapters, we will derive measures for the degree of separation between two motions and develop the conditions for motion-background, translational-motion and planar-motion segmentation.

Chapter 4

Analysis of Motion-Background Segmentation

In most computer-vision applications, feature or image points extracted from images of a dynamic scene are associated with either moving objects or the background. In this categorisation, the background points have zero or negligible motion as they are generally extracted from static objects or objects located very far away from the camera.

This chapter studies the feasibility of detection and segmentation of an unknown motion from its static background using the fundamental matrix motion model. First, the separability of a pure translation from static background is theoretically analysed in section 4.1. The analysis shows that a pure translation is not separable from its static background using the fundamental matrix. Section 4.2 then proposes a set of sufficient conditions for motion-background segmentation based on a quantitative measure for the degree of separation in terms of the rotation angle of the target

motion. The results of the experiments using real-image data designed to demonstrate the usability of the proposed conditions are presented and discussed in section 4.3. Finally, section 4.4 concludes the chapter.

4.1 Non-separability of a pure translation

To analyse the feasibility of a motion-background segmentation, a dynamic scene containing two rigid 3D-objects having distinct motions, denoted as motion-a and motion-b, is considered. The focus is on a scene viewed by a static and uncalibrated camera (as introduced in section 3.1). Motion-a is parameterised by a rotation θ_a around the camera optical axis and followed by a non-zero translation T_a, while the parameters of motion-b are set to zero, i.e. $\theta_b = 0$ and $T_b = 0$, to represent a static object or background.

The analysis aims to prove that a pure translational motion is not separable from static points in a motion segmentation using the fundamental matrix. In other words, we aim to segment points having a pure translational motion from static points associated with the background. Firstly, the analysis focuses on a case of motion-a restricted to a 2D translational-motion, i.e. θ_a is set to zero and T_a is confined on a plane perpendicular to the camera optical axis (Z axis); this motion is denoted by the symbol $T_{a2D} = [T_{xa}\ T_{ya}\ 0]^\top$. The 2D translational-motion segmentation analysis provides the necessary foundation for a more general case, where motion-a is an arbitrary 3D translation $T_a = [T_{xa}\ T_{ya}\ T_{za}]^\top$ (including a component along the camera optical axis). Without loss of generality, in terms of robust estimation, the image points having T_{a2D} or T_a extracted from the images of the scene are considered *inliers*,

which are to be separated from the points associated with the static background — the *outliers*.

In this analysis, we assumed that an estimator has provided an accurate estimate of the fundamental matrix of a 2D translation T_{a2D}

$$F_{Ta2D} = \frac{1}{f}\begin{bmatrix} 0 & 0 & T_{ya} \\ 0 & 0 & -T_{xa} \\ -T_{ya} & T_{xa} & 0 \end{bmatrix}, \tag{4.1}$$

provided by equation (2.5) [3, 39, 124] with camera matrix according to A in (2.2) including focal length f and the image principal point $[P_x \ P_y]$. The relationship between the noise-free points, denoted by the underlined symbols, having a 2D translation T_{a2D} in two images are

$$\begin{aligned} \underline{x}_{1i} &= \frac{fX_i}{Z_i} + P_x, \quad \underline{x}_{2i} = \underline{x}_{1i} + \frac{fT_{xa}}{Z_i} + P_x, \\ \underline{y}_{1i} &= \frac{fY_i}{Z_i} + P_y \quad \text{and} \quad \underline{y}_{2i} = \underline{y}_{1i} + \frac{fT_{ya}}{Z_i} + P_y, \end{aligned} \tag{4.2}$$

where the symbols $[X_i \ Y_i \ Z_i]$ refer to the coordinates of the point in 3D-space. While the noise-free points associated with the static background have the following relationship in two images:

$$\begin{aligned} \underline{x}_{1i} &= \frac{fX_i}{Z_i} + P_x, \quad \underline{x}_{2i} = \underline{x}_{1i}, \\ \underline{y}_{1i} &= \frac{fY_i}{Z_i} + P_y \quad \text{and} \quad \underline{y}_{2i} = \underline{y}_{1i}. \end{aligned} \tag{4.3}$$

All measured points are assumed to be contaminated by measurement noise e; having a Gaussian distribution and being independently and identically distributed (i.i.d)

$$\begin{aligned} x_{1i} &= \underline{x}_{1i} + e^1_{ix}, \qquad y_{1i} = \underline{y}_{1i} + e^1_{iy}, \\ x_{2i} &= \underline{x}_{2i} + e^2_{ix}, \qquad \text{and} \qquad y_{2i} = \underline{y}_{2i} + e^2_{iy}, \end{aligned} \tag{4.4}$$

where $[e^1_{ix}\ e^1_{iy}\ e^2_{ix}\ e^2_{iy}]^\top \sim N(0, \sigma_n^2 I_4)$ in which σ_n is the unknown scale of noise and I_4 is a 4×4 identity matrix. The Sampson distances of all points are computed using equation (3.4) with the substitution of the fundamental matrix of the target translation T_{a2D} in (4.1) and the noise contamination in (4.4). This yields:

$$d_i = \frac{T_{ya}(\underline{x}_{2i} + e^2_{ix} - \underline{x}_{1i} - e^1_{ix}) + T_{xa}(\underline{y}_{1i} + e^1_{iy} - \underline{y}_{2i} - e^2_{iy})}{\sqrt{2(T^2_{ya} + T^2_{xa})}}. \tag{4.5}$$

For points having T_{a2D}, the above expression without noise terms are equal to zero according to equation (2.3) since the true F_{Ta2D} is used to compute the distances d_i. Thus, for points having T_{a2D}, equation (4.5) can be simplified to:

$$d_i = \frac{T_{ya}(e^2_{ix} - e^1_{ix}) + T_{xa}(e^1_{iy} - e^2_{iy})}{\sqrt{2(T^2_{ya} + T^2_{xa})}}. \tag{4.6}$$

The distances in equation (4.6) are a linear combination of the i.i.d. noise samples e. Therefore, they are also normally distributed with zero mean and variance σ_n^2 as the numerator and denominator cancel each other.

The distances associated with static points can also be calculated using equation (4.5), the world-to-image points relationship in (4.3) and the measurement noise assumption in (4.4), which results in the same equation as that given in (4.6). Thus, the distances of the static background are also distributed according to N(0,σ_n^2). Hence, the distributions of distances associated with points having the target translation T_{a2D} and the static background will be the same.

The success of motion segmentation is determined by looking at the relative sizes of *inlier* and *outlier* distances. In order for the points having T_{a2D} to be successfully segmented, the smallest distance associated with the static background should be sufficiently larger that the biggest distances associated with points having T_{a2D} (note

that the distances are calculated using the fundamental matrix associated with the moving object). Since the distributions of distances associated with points having T_{a2D} and static background are the same, theoretically, the points having T_{a2D} cannot be segmented from the static background if the fundamental matrix motion model is used.

The segmentation of 3D translation T_a with $T_{za} \neq 0$ form static background is mathematically intractable and too complex to be derived theoretically. However, the results of our Monte Carlo experiments in section 4.2 verify that the Sampson distances of the points having 3D translation and static background are also normally distributed with zero mean and similar variances. Therefore they are also not separable from the static background.

4.2 Conditions for motion-background segmentation

The non-separability of a pure translation from its static background when using the fundamental matrix as motion model implies that the separability of a motion from its background depends on its rotational part. Therefore we aimed to determine the sufficient conditions in terms of the minimum rotation angle for successful motion-background segmentation via Monte Carlo experiments using synthetic images. The correctness of these conditions was verified by studying the variance of the experimental results.

The Monte Carlo experiments consisted of two main parts: the first part was to

verify the non-separability of a pure 3D translation from its static background and the second part was to determine the minimum rotation angle for correct and successful segmentation. The Monte Carlo experimental setup was designed to be identical with the earlier analysis in section 4.1.

In each experiment, 2000 randomly generated points $M_{1i} = [X_{1i}\ Y_{1i}\ Z_{1i}]^\top$ in a world-coordinate system were moved to position $M_{2i} = [X_{2i}\ Y_{2i}\ Z_{2i}]^\top$ according to motion-a. The points M_{1i} to M_{2i} were viewed by a static camera A_1, with a focal length of 703 pixels and image size of 512×512 with the principal point at the center of the images i.e.

$$A_1 = \begin{bmatrix} 703 & 0 & 256 \\ 0 & 703 & 256 \\ 0 & 0 & 1 \end{bmatrix}, \tag{4.7}$$

representing a typical camera with a field of view around 40°. The symbols $\underline{m}_{1i} = [\underline{x}_{1i}\ \underline{y}_{1i}]^\top$ and $\underline{m}_{2i} = [\underline{x}_{2i}\ \underline{y}_{2i}]^\top$ denoted the corresponding points on two images.

The static points representing the background were randomly added to both images based on the magnitude of the intended *inlier* ratio ϵ shown in equation (3.2). All randomly generated image points were perturbed with a Gaussian noise with zero mean and a standard deviation of σ_n. We aimed to segment points having motion-a from the static points using a robust estimator and the fundamental matrix motion model.

The true fundamental matrix of points having motion-a F_a was calculated using equation (2.5) by substitution of the known motion parameters and camera matrix A_1 in (4.7). The residuals in terms of Sampson distances d_i^2 were computed for all points (both moving and static points) using equation (3.4) and the true F_a.

Motion segmentation according to the steps in MSSE [6] (shown in equation 3.6) was performed to identify and segment points having motion-a from the static background.

The segmentation performance was measured by the ratio (ζ) of the number of segmented points having motion-a over the true number of points having motion-a. The translational parameters T_a were randomly selected and every set of the experiments was repeated 1000 times for incremental θ_z (from 0° to 70°) and for various values of *inlier* ratios (from $\epsilon = 30\%$ to 80%), measurement noise ($\sigma_n = 0.25$ to 2) and camera parameters according to A_1 in equation (4.7) and randomly selected values of focal length in A_2 and A_3:

$$A_2 = \begin{bmatrix} 492.1 & 0 & 256 \\ 0 & 492.1 & 256 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{and} \quad A_3 = \begin{bmatrix} 527.3 & 0 & 256 \\ 0 & 597.6 & 256 \\ 0 & 0 & 1 \end{bmatrix}. \tag{4.8}$$

The selection of camera matrices A_2 and A_3 in (4.8) represents both homogenous (equal focal lengths in X and Y directions) and non-homogenous changes to the camera matrix A_1.

In order to analyse the performance and consistency of the segmentation, the statistical mean and standard deviation of 1000 ζs (denoted by $\bar{\zeta}$ and σ_ζ) were calculated and recorded. The segmentations of motion-a were considered to be correct and consistent if $\bar{\zeta} \approx 1$ and $\sigma_\zeta \leq 0.01$ throughout the 1000 ζs. The pseudo code of the Monte Carlo experiments is given in figure 4.1.

The first part of the experiments was designed to verify the earlier statement on the non-separability of pure translational motions from static background in section 4.1. In the experiment, motion segmentation was performed to identify the points having either a random 2D translation (T_{a2D} with a zero translational component

along the camera optical axis $T_{za} = 0$) or 3D translation (T_a with $T_{za} \neq 0$) from static points in the background. Specifically, two thousand randomly selected image points, having either translation T_{a2D} and T_a, were mixed with random static points while the *inlier* ratio was at 80% and the standard deviation of Gaussian noise σ_n was equal to one. Motion-background segmentation was performed and the segmentation performance ζ and two types of scale estimates were calculated; these were the scale estimates associated with the ground-truth moving points (the *inlier* scale denoted by S) and the scale estimate given by all data points (the total scale denoted by Γ).

Repeat (*inlier* ratio $\epsilon = 30\%$ to 80%) and (noise level $\sigma_n = 0.25$ to 2).
Repeat (rotation angle $\theta_z = 0°$ to $70°$) and (camera matrix $A = A_1$ to A_3).

i. Repeat ($j = 1$ to 1000).
   1. Generate a random translation $T_a = [T_{xa}\ T_{ya}\ T_{za}]^\top$ (between $\pm$ 0.5m).
   2. Generate 2000 random pairs of moving points according to θ_z and T_a.
   3. Generate random pairs of static points based on ϵ in (3.2).
   4. Project all points on two images using a camera matrix A.
   5. Perturb all points with Gaussian noise $N(0, \sigma_n^2)$.
   6. Calculate the true F of the moving points.
   7. Calculate the Sampson distances using the true F.
   8. Sort the square of Sampson distances d_i^2 and perform segmentation using MSSE.
   9. Record the scale estimate given by the moving points (*inlier* scale S) and by all image points (total scale Γ), separately.
   10. Record the ratio of the number of segmented moving points over the true number of moving points ζ.

ii. End.

iii. Calculate and record the mean and standard deviation of the 1000 Ss, Γs and ζs.

End, End.

Figure 4.1: Pseudocode of the Monte Carlo experiments for the analysis of motion-background segmentation.

For two instances of the random image points having translations T_{a2D} and T_a, we have recorded the corresponding values of S, Γ and segmentation performance ζ in table 4.1 and plotted the histogram of distances associated with all image points (moving and static points) in figure 4.2. In both cases, the segmentations were not successful since a number of static points were segmented as points having either T_{a2D} or T_a, indicated by values of ζ larger than one in table 4.1. This is because the populations of distances associated with moving points could not be distinguished from the distances associated with static points, as shown in figure 4.2. In addition, the values of S and Γ were very close to each other, i.e. very close to one or σ_n, showing that the distances associated with either moving or static points were overlapping each other. Hence, the moving points were not distinguishable from the static background.

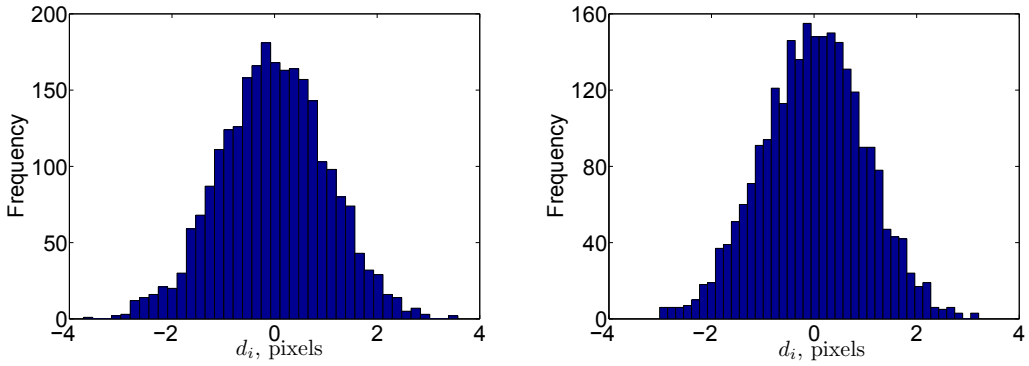
To examine the effect of *inlier* ratio to the non-separability of a pure translation from static points, the experiment was repeated 1000 times while varying ϵ from 30% to 80% with randomly selected 3D translation T_a. The standard deviation of the measurement noise was maintained at one. The statistical mean and standard deviation of *inlier* scales S and total scales Γ, represented by the symbols $\bar{S}$, σ_S, $\bar{\Gamma}$ and σ_Γ respectively, were calculated throughout the 1000 iterations of each experiment. Table 4.2 summarises the values of $\bar{S}$, σ_S, $\bar{\Gamma}$ and σ_Γ for all *inlier* ratios.

The values of $\bar{S}$ and $\bar{\Gamma}$ are very close to $\sigma_n = 1$, as observed in table 4.2. In addition, the values of σ_S and σ_Γ in table 4.2 are very small and close to zero, indicating that the values of *inlier* and total scale (S and Γ) were consistent throughout the experiments. These observations signify that a pure translational motion is not separable from the static points, regardless of translational parameters and location of points (as there were randomly selected throughout the experiments) and *inlier*

ratios. These results, which are consistent with our earlier analysis in section 4.1, show that a pure translational motion cannot be segmented from static points using the fundamental matrix motion model.

Table 4.1: Results for motion-background segmentation involving pure translations when $\epsilon = 80\%$ and $\sigma_n = 1$.

	$[T_{xa}\ T_{ya}\ T_{za}]^\top$ cm	Inlier scale S	Total scale Γ	ζ
T_{a2D}	[-9.79 -2.67 0]	1.0066	1.0082	1.22
T_a	[-5.65 -2.36 -9.16]	1.0012	1.0075	1.23



(a) 2D translation, T_{a2D}

(b) 3D translation, T_a

Figure 4.2: Distribution of Sampson distances for all image points in motion-background segmentation involving pure translations.

Table 4.2: *Inlier* scale and total scale for various *inlier* ratio ϵ.

ϵ, %	static points	$\bar{S}$	σ_S	$\bar{\Gamma}$	σ_Γ
80	500	1.0001	0.0156	0.9999	0.0139
70	857	1.0003	0.0161	1.0001	0.0132
60	1333	0.9994	0.0160	0.9995	0.0124
50	2000	0.9994	0.0154	0.9997	0.0113
40	3000	0.9998	0.0159	1.0000	0.0104
30	4667	1.0009	0.0164	1.0002	0.0093

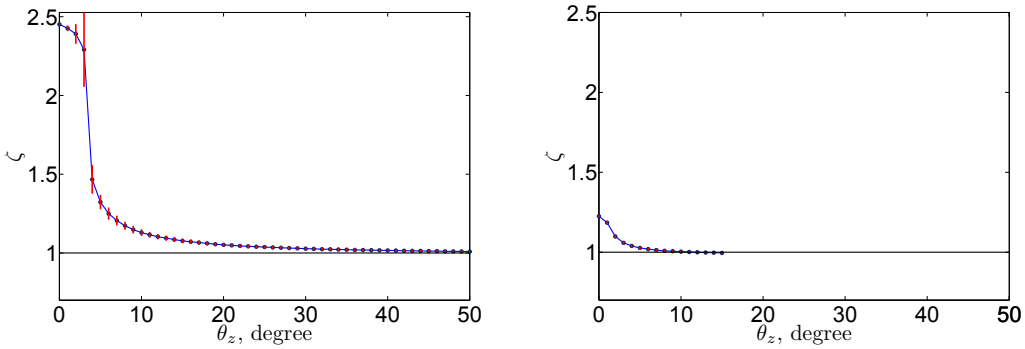
The absence of success in motion-background segmentation when the motion is a pure translation implies that the feasibility of motion-background segmentation depends on the rotational part of the motion. Hence, the second part of the experiments was designed to show that the rotation angle of a motion could be used as a measure for the degree of separation between a motion from static points associated with the background. The experiments also aimed to determine the condition for correct and successful motion-background segmentation in terms of minimum rotation angle.

In the experiments, we examined the performance of motion-background segmentation of motion-a, including rotation θ_z around the camera optical axis and followed by a random 3D translation. Several scene and motion parameters were varied in the experiments, namely, the rotation angle θ_z (from 0° to 70°), the *inlier* ratio ϵ (30% to 80%) and the noise levels σ_n (0.25 to 2). In addition, to examine the effect of camera variation on the segmentation performance, the experiments were also repeated for camera matrices A_2 and A_3 in (4.8) for variation of A_1 in (4.7) (the focal length of camera matrices A_2 and A_3 were randomly selected to represent both homogenous

(equal focal lengths in X and Y directions) and non-homogenous changes to camera matrix A_1). The points having motion-a and static points from the background were randomly selected and motion segmentation was repeated 1000 times for each value of rotation angle θ_z, *inlier* ratio ϵ, level of noise σ_n and using camera matrices (A_1 to A_3). To analyse the performance and consistency of the segmentation, the statistical mean and standard deviation of segmentation performance ($\bar{\zeta}$ and σ_ζ) were also calculated and recorded.

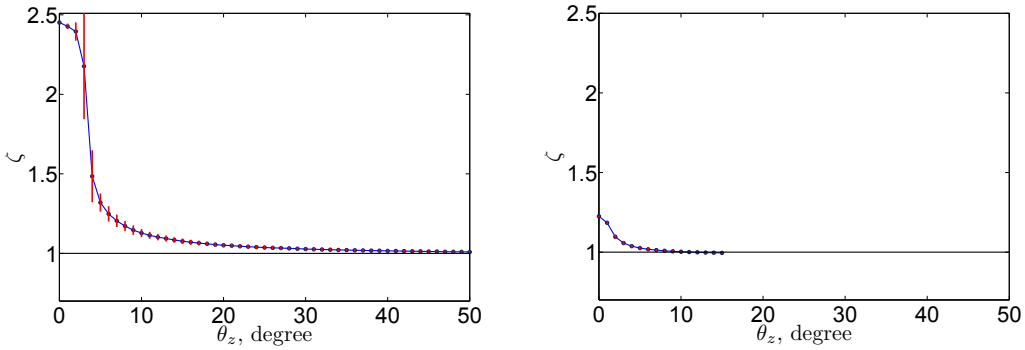
Figures 4.3 and 4.4 show $\bar{\zeta}$ and σ_ζ versus θ_z when σ_n=1 and ϵ=40% or 80% for camera matrices A_1 to A_3. It can be observed from figure 4.3(b) that, for small rotations, i.e. θ less than around 11° when $\epsilon = 80\%$, ζ is larger than one, indicating that some static points were segmented as points having motion-a. In such cases, the inaccurate dichotomy between moving and background points resulted in an incorrect motion-estimation and segmentation. As the value of θ_z increased from 0° to 50° (in figures 4.3(a) and 4.3(b)), the values of $\bar{\zeta}$ and σ_ζ reduced to one and zero, respectively. This indicated perfect and consistent segmentation (ζ around one), which occurred when the value of θ_z was larger than around 35° when ϵ was 40%, and 11° when ϵ was 80%. Good consistency of segmentation, i.e. $\sigma_\zeta \approx 0$ when $\theta_z > 35°$ and 11° for $\epsilon = 40\%$ and 80% in figures 4.3(a) and 4.3(b), also signifies that locations (the coordinates) of image points and translational parameters did not affect the segmentation performance since they were randomly selected throughout the experiments. These results show that the rotation angle could be used as a measure for the degree of separation in motion-background segmentation problems. In addition, the identical results seen while comparing figure 4.3(a) with 4.4(a) and figure 4.3(b) with 4.4(b), show that variations of camera parameters do not affect the performance of

motion-background segmentation.



(a) $\epsilon = 40\%$ (b) $\epsilon = 80\%$

Figure 4.3: $\bar{\zeta}$ and σ_ζ vs θ_z using camera matrix A_1.



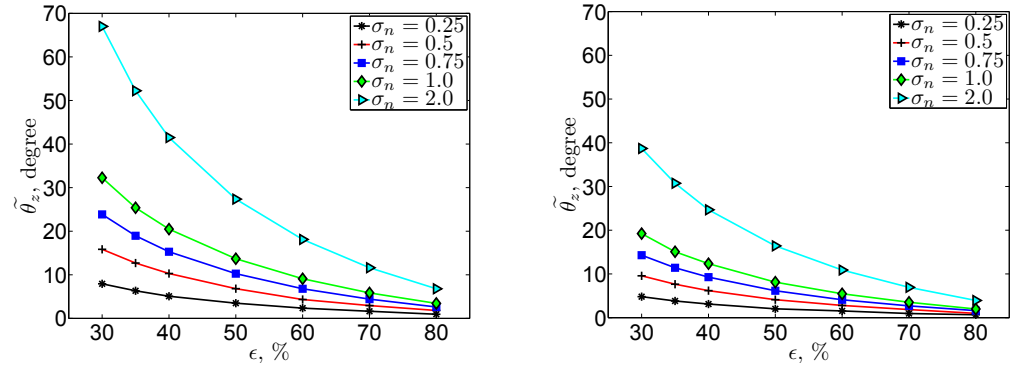
(a) ϵ =40% (b) ϵ =80%

Figure 4.4: $\bar{\zeta}$ and σ_ζ vs θ_z using camera matrix A_2 in (a) and A_3 in (b).

In the target applications, such as in traffic or video surveillance system, an accuracy of 90% or higher is usually considered a good accuracy. Thus, throughout the experiments we considered two accuracy thresholds; 95% and 90% accuracies, i.e.

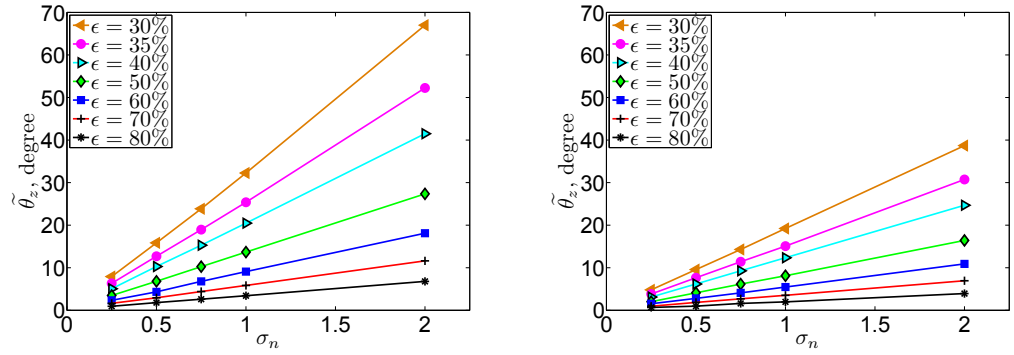
with acceptable 5% or 10% errors, respectively. Specifically, the motion-background segmentation were considered successful and consistent when $\bar{\zeta_1} = 1.05$ and $\bar{\zeta_2} = 1.10$, both with $\sigma_\zeta \approx 0.01$. Using these thresholds meant that it was accepted that 5% or 10% of background points with small distances were to be segmented as moving points. In addition, the segmentation results were expected to be consistent since the values of σ_ζ associated with both $\bar{\zeta_1}$ and $\bar{\zeta_2}$ were very close to zero. The conditions for successful motion-background segmentation, in terms of the minimum rotation angles (denoted by $\widetilde{\theta_z}$), were then interpolated for both thresholds $\bar{\zeta_1}$ and $\bar{\zeta_2}$ with $\sigma_\zeta \approx 0.01$ from the plots of ζ versus θ_z for all values of *inlier* ratio ϵ and noise levels σ_n.

A broad picture of the condition for motion-background segmentation (the required $\widetilde{\theta_z}$) for various *inlier* ratio and noise levels is shown in figures 4.5 and 4.6. These results show that the motion-background segmentation becomes more challenging and difficult, as indicated by the larger values of required $\widetilde{\theta_z}$, when there are many points associated with the static background (small value of ϵ) and/or high level of noise (large value of σ_n). When the value of the standard deviation of measurement noise was high, the maximum residual associated with moving points became large. Consequently, more static points with smaller residuals were mixed with the residuals of moving points, and more background points were likely to be segmented as moving points. As such, high value of measurement noise would result in a more difficult motion-background segmentation problem thus, a larger magnitude of $\widetilde{\theta_z}$ was required to produce sufficiently distinct residuals to differentiate between moving and static points for successful segmentation.



(a) ζ_1=1.05 (b) ζ_2=1.10

Figure 4.5: $\widetilde{\theta}_z$ vs ϵ for various σ_n.



(a) ζ_1 =1.05 (b) ζ_2 =1.10

Figure 4.6: $\widetilde{\theta}_z$ vs σ_n for various ϵ.

The experimental results show that the rotation angle could be used as a measure for the degree of separation in motion-background segmentation problems. A set of sufficient conditions to guarantee successful motion-background segmentation was proposed and these conditions were shown to be independent of the translational parameters of the motion, the location of points in image plane and the camera

parameters. If these conditions are satisfied, i.e. θ_z larger than or equal to $\widetilde{\theta}_z$, the segmentation is guaranteed to achieve the targeted accuracies ($\bar{\zeta_1} = 1.05$ and $\bar{\zeta_2} = 1.10$). The relevance of the proposed conditions and its usage in motion-segmentation applications will be examined via experiments using real-image data in section 4.3.

4.3 Experiments using real images

The relevance of the proposed conditions for motion-background segmentation and its applicability were examined via experiments using real-image data. We have again considered a scene containing a moving object and a static object in the background. The experimental aim was to identify and segment the points having motion from the mixture of moving and static points.

The experiments focus on investigating the theoretical limit of motion-background segmentation and how imperfect estimate of the fundamental matrix would affects the conditions for segmentation is beyond the scope of this work. In practice, the fundamental matrix can be accurately estimated using a number of robust methods [3, 101, 124] and the *gross outliers* can be removed by applying a robust estimator as part of the motion segmentation process [123]. The issues of estimation including estimating the fundamental matrix in terms of both the feasibility and the accuracy, have already been thoroughly analysed [42, 44]. Thus, in our experiments using real-image data — identical to our earlier theoretical analysis and Monte Carlo experiments — we assumed that an accurate estimate of the fundamental matrix of the motion was provided by a robust estimator and there were no mismatches (*gross outliers*) in the image data. As such we calculated the fundamental matrix of the motion using

equation (2.5) and manually removed occasional *gross outliers* in the data. These assumptions needed to be taken in order to eliminate the effect of potential errors from the estimation of the fundamental matrix and the presence of *gross outliers*, to the experimental results.

The experiments consisted of three steps:

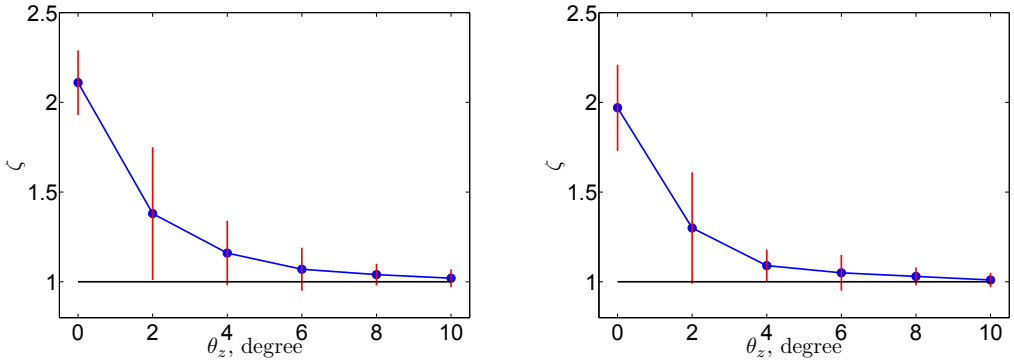
1. Camera calibration and image acquisition: The dynamic scene in the experiments consisted of a specially designed and fabricated 3D object having motion-a and a book to represent the static background. The camera was calibrated using a publicly available camera-calibration toolbox [16]. The 3D object was moved according to motion-a parameterised by rotation θ_z and followed by translation T_a. The values of θ_z were from $\theta_z = 0^\circ$ to 10°, with 2° in each increment, and each θ_z was followed by twenty sets of distinct 3D translation T_a. The image of the scene was taken both before and after each motion-a (pair of θ_z and T_a). In total, 120 pairs of images were used in the experiments.

2. Image-data preparation: Image distortion was reduced from all images using radial and tangential distortion models suggested by a camera-calibration toolbox [16]. All corresponding image points associated with either moving objects or static backgrounds were extracted using a publicly available implementation of the Scale-Invariant Feature Transform (SIFT) algorithm [58, 56] and occasional incorrect matches were manually removed. The values of *inlier* ratio ϵ in each pair of images was varied from 35% to 80% by removing some of the points associated with the background.

3. Segmentation analysis: The fundamental matrix for motion-a and the distances d_i associated with all points in each pair of images were calculated using equations (2.5) and (3.4) with the known values of θ_z, T_a and the camera matrix from camera calibration. Motion segmentation was performed to recover all points having motion-a from each pair of images using the segmentation steps according to the Modified Selective Statistical Estimator in (3.6) [6]. The segmentation performance was measured by calculating the value of ζ, i.e. the ratio of the segmented points over the number of ground-truth points having motion-a. The standard deviation of measurement noise σ_n throughout the experiments was estimated from equation (3.5).

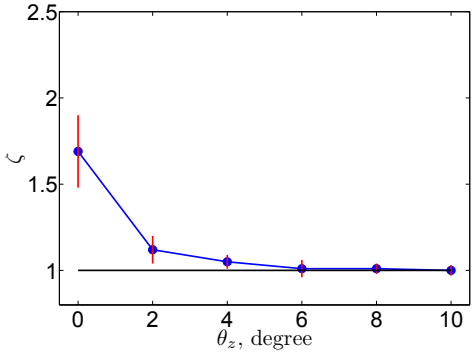
Figure 4.7 shows the mean and standard deviation of ζ ($\bar{\zeta}$ and σ_ζ) versus θ_z (there are twenty values of ζs for each θ_z) for various *inlier* ratios. It can be observed from figure 4.7 that, when $\theta_z = 0°$, where motion-a is a pure translational motion, the segmentation performance $\bar{\zeta}$ is greater than one, indicating unsuccessful segmentations. When the values of θ_z associated with motion-a are increased to 10°, the values of $\bar{\zeta}$ and σ_ζ reduced to around one and zero, respectively. This signifies improving and successful motion-background segmentations. These results imply two main points; motion-background segmentation is not possible for a pure translational motion and the magnitude of rotation angle θ_z could be used as a measure of separability between moving and static points.

The relevance of the derived condition for segmentation from our Monte Carlo experiments (in figures 4.5 and 4.6) was examined by comparing them with the results from experiments using real-image data — where the required θ_z was interpolated

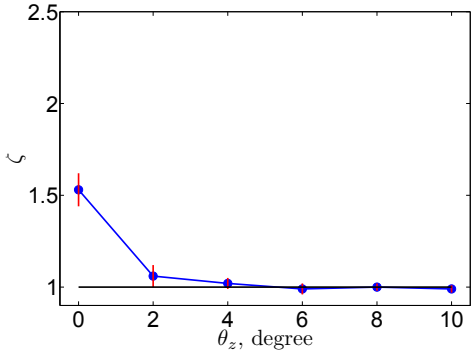


(a) ϵ =35%

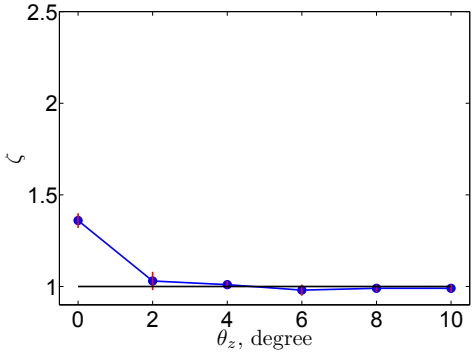
(b) ϵ=40%



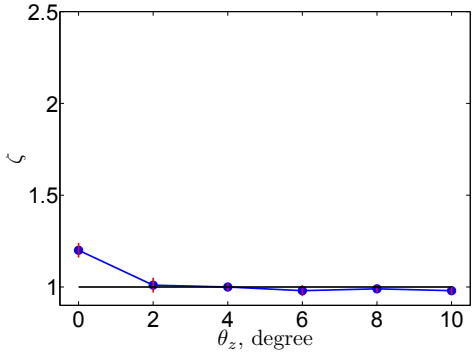
(c) ϵ =50%



(d) ϵ =60%

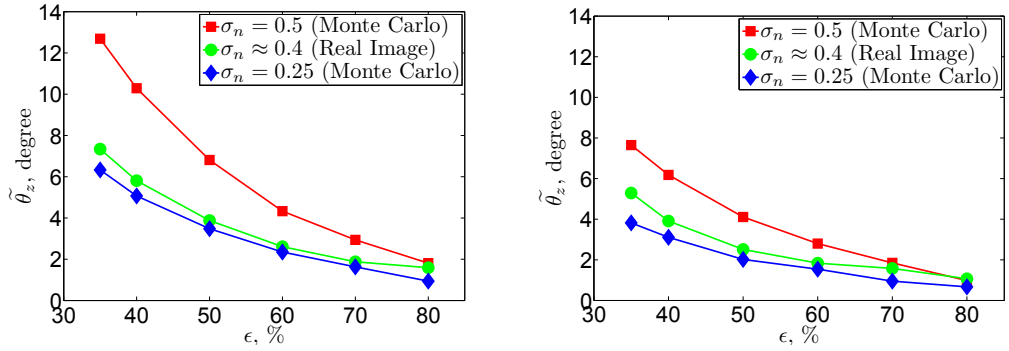


(e) ϵ=70%



(f) ϵ =80%

Figure 4.7: Mean and standard deviation of ζ vs θ_z for various ϵ.



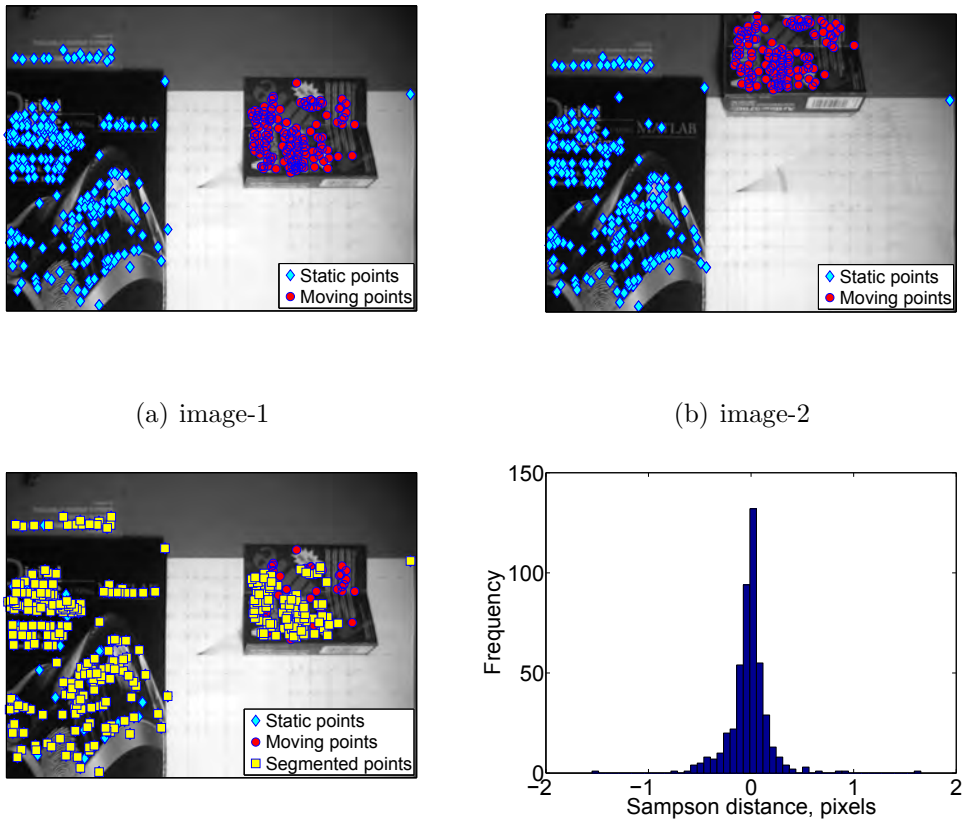
(a) $\zeta_1 = 1.05$ (b) $\zeta_2 = 1.10$

Figure 4.8: $\widetilde{\theta_z}$ vs ϵ for Monte Carlo and real-image experiments.

at $\bar{\zeta} \approx 1$ from figure 4.7. Figure 4.8 shows the conditions for motion-background segmentation from both synthetic and real-image data with the standard deviation of noise estimated around 0.4 pixels (from equation (3.5)) throughout the experiments. Similar trends in $\widetilde{\theta_z}$ versus *inlier* ratios can be observed in figure 4.8 for results from both experiments using synthetic and real-image data.

To provide an insight into the capability of the derived conditions (in figures 4.5 and 4.6) to predict the outcomes of motion-background segmentations, we have examined three cases where motion-a was a pure translation ($\theta_z = 0°$) and motion-a having either θ_z less than or greater than the required threshold. The selected threshold is $\widetilde{\theta_z} \approx 6°$ (extrapolated from figure 4.6(a)) to guarantee successful segmentation with accuracy of $\zeta_1 = 1.05$ when *inlier* ratio is 35%. It was predicted that the segmentation would be successful when $\theta_z \geq \widetilde{\theta_z}$ ($\theta_z = 8°$) and fail when $\theta_z = 0°$ and $\theta_z \leq \widetilde{\theta_z}$ ($\theta_z = 4°$). Figures 4.9(c) and 4.10(c) show that when motion-a is a pure translation ($\theta_z = 0°$) or motion-a includes $\theta_z \leq \widetilde{\theta_z}$, the segmentations are unsuccessful, indicated by the values of ζ greater than the expected accuracy, i.e. 1.05% (ζ around 2.40 and

1.12 when $\theta_z = 0°$ and $\theta_z \leq \widetilde{\theta}_z$). This is because the distances d_i associated with moving and static points are overlapping and cannot be easily distinguished from each other, as shown in figures 4.9(d) and 4.10(d).



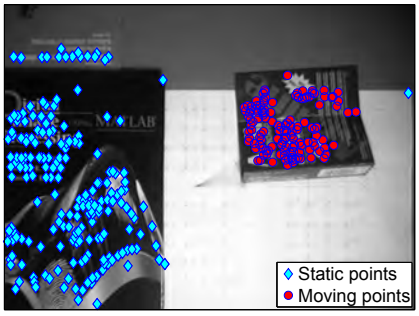
(a) image-1

(b) image-2

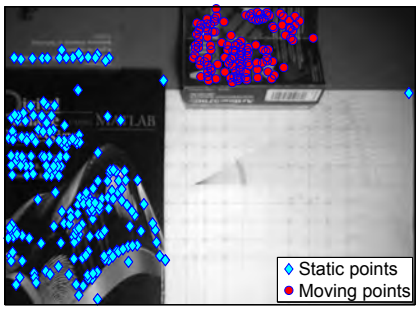
(c) $\zeta = 2.40$

(d)

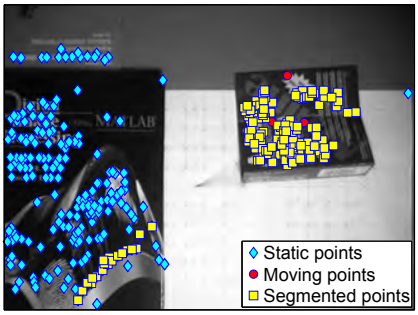
Figure 4.9: The ground-truth when motion-a is a pure translation $T_a = [-59 \ -82 \ -39]^\top$mm ($\theta_z = 0°$) and $\epsilon = 35\%$ in (a) and (b). The segmented points having motion-a in (c) and the histogram of d_i for all points in (d).



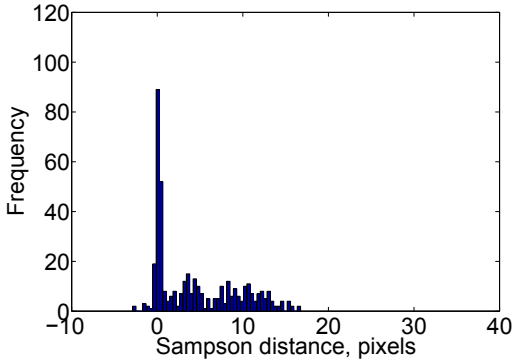
(a) image-1



(b) image-2

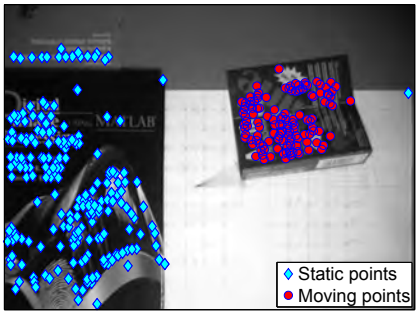


(c) $\zeta = 1.12$

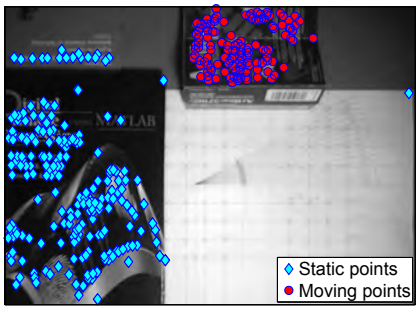


(d)

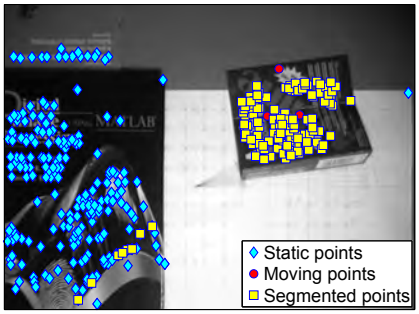
Figure 4.10: The ground-truth when motion-a is parameterised by $\theta_z = 4°$ and $T_a = [-59 \;\; -82 \;\; -39]^\top$mm with $\epsilon = 35\%$ in (a) and (b). The segmented points having motion-a in (c) and the histogram of d_i for all points in (d).



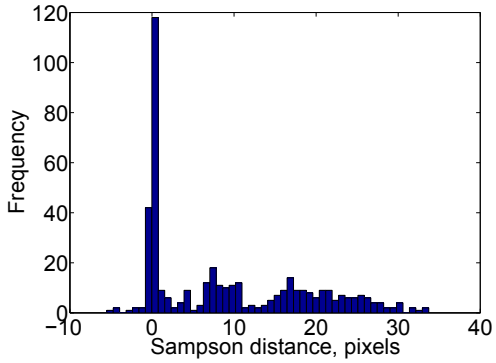
(a) image-1



(b) image-2



(c) $\zeta = 1.03$



(d)

Figure 4.11: The ground-truth when motion-a is parameterised by $\theta_z = 8°$ and $T_a = [-59 \;\; -82 \;\; -39]^\top$mm with $\epsilon = 35\%$ in (a) and (b). The segmented points having motion-a in (c) and the histogram of d_i for all points in (d).

As predicted when motion-a consists of θ_z larger than the required $\widetilde{\theta}_z$, the segmentation was considered successful, since the value ζ was around 1.03, which was better than the expected accuracy of 1.05% as seen in figure 4.11(c). This is because the distances d_i associated with static points were more spread and could be distinguished from d_i of moving points by a robust estimator as shown in figure 4.11(d).

These results show that the derived conditions from the Monte Carlo experiments are very relevant to the problem encountered in real-world applications. They also demonstrate the capability of the proposed conditions to guarantee successful motion-background segmentation with accuracies of 105% or 110%.

4.4 Conclusion

The theoretical analysis and the experimental results in this chapter revealed two main points. Firstly, a pure translational motion is not separable from static background using fundamental matrix motion model. Secondly, the success of motion-background segmentation using fundamental matrix depends on the rotation angle of that particular motion. Sufficient conditions for motion-background segmentation were proposed, in terms of minimum rotation angle via extensive experiments using synthetic images. These conditions indicated that the segmentation became more challenging in a scene including large number of static points and high level of measurement noise. Experiments using real-image data showed that the conditions for segmentation were very relevant to real motion-background segmentation problems. In practice, the conditions are capable of predicting the outcome of motion-background segmentation using obtainable amount of rotation associated with a particular motion.

Chapter 5

Analysis of Translational-Motion Segmentation

A number of computer-vision applications, such as traffic surveillance system, require the recovery of a pure translational motion. To develop some guidelines for the design of these systems, the feasibility of motion segmentation involving translational 3D objects using the translational fundamental matrix is analysed in this chapter. The focus is on translations of rigid 3D objects viewed by an uncalibrated camera. The feasibility of segmentation is expressed as a set of conditions for successful segmentation.

The analysis starts with the modelling of a dynamic scene including multiple translating objects in section 5.1. We then derive a quantifiable measure of separation between two 2D translations (translations restricted on a plane perpendicular to the camera optical axis) in section 5.2. The 2D-derivation provides the necessary foundation for finding the requirements for successful segmentation of an arbitrary

3D translational-motion (translation including components along the camera optical axis) in section 5.3. The conditions for successful segmentation are proposed via theoretical analysis and extensive Monte Carlo experiments using synthetic images. Section 5.4 details the experiments using real images designed to demonstrate the application of the proposed conditions to correctly predict the outcome of motion segmentations. Section 5.5 concludes the chapter.

5.1 Dynamic-scene representation

The analysis considers a dynamic scene including two rigid 3D-objects moved according to two distinct translations denoted by T_a and T_b where $T_a = [T_{xa}\ T_{ya}\ T_{za}]^\top$ and $T_b = [T_{xb}\ T_{yb}\ T_{zb}]^\top$. The background or static features in the scene are not taken into account as their effect has been considered in chapter 4.

Consider a point in 3D-space with coordinates $[X_i\ Y_i\ Z_i]^\top$ and denote its corresponding point in the image plane by $\underline{m}_{1i} = [\underline{x}_{1i}\ \underline{y}_{1i}]^\top$ which moves to $\underline{m}_{2i} = [\underline{x}_{2i}\ \underline{y}_{2i}]^\top$ after a 3D translation. The relationships between the corresponding image-world coordinate points viewed by a perspective camera A (as shown in equation (2.2)) are

$$\begin{aligned} \underline{x}_{1i} &= \frac{fX_i}{Z_i} + P_x, \qquad \underline{x}_{2i} = \frac{f(X_i + T_x)}{Z_i + T_z} + P_x, \\ \underline{y}_{1i} &= \frac{fY_i}{Z_i} + P_y, \qquad \underline{y}_{2i} = \frac{f(Y_i + T_y)}{Z_i + T_z} + P_y, \end{aligned} \tag{5.1}$$

where the symbols T_x, T_y and T_z in (5.1) represent the translation parameters. All image points are assumed to be contaminated by independently and identically dis-

tributed (i.i.d) measurement noise with Gaussian distribution

$$\begin{aligned} x_{1i} &= \underline{x}_{1i} + e^1_{ix}, & y_{1i} &= \underline{y}_{1i} + e^1_{iy}, \\ x_{2i} &= \underline{x}_{2i} + e^2_{ix}, \quad \text{and} & y_{2i} &= \underline{y}_{2i} + e^2_{iy}, \end{aligned} \tag{5.2}$$

where $[e^1_{ix}\ e^1_{iy}\ e^2_{ix}\ e^2_{iy}]^\top \sim N(0, \sigma_n^2 I_4)$ in which σ_n is the unknown scale of noise and I_4 is the 4×4 identity matrix. The underlined variables denote the true noise-free locations of the points in the image plane. In this analysis, translation T_a is considered as the target motion and T_b is the unwanted or the other translation. Without loss of generality, in the context of robust estimation, the matching points having T_a are assumed to be *inliers* aimed to be segmented from the matching points having T_b, which are considered as *outliers*.

The fundamental matrix of translation T_a parameterised by T_{xa} T_{ya} and T_{za} is computed using equation (2.5) [3, 39, 124]:

$$F_{Ta} = \frac{1}{f^2} \begin{bmatrix} 0 & -T_{za} & T_{ya}f + T_{za}P_y \\ T_{za} & 0 & -T_{xa}f - T_{za}P_x \\ -T_{ya}f - T_{za}P_y & T_{xa}f + T_{za}P_x & 0 \end{bmatrix}. \tag{5.3}$$

Similar to chapter 4, motion segmentation is performed using the MSSE [6] and the square of Sampson distances (shown in equation (3.4)) [101, 119] as residuals.

The analysis of the feasibility of segmentation for arbitrary 3D translational-motion is relatively complicated. To simplify the presentation of the analysis, we first concentrate on the case of 2D translation in a plane perpendicular to the camera optical axis. These results are then used as the basis for presenting the analysis of arbitrary 3D translational-motion with component along the camera optical axis (T_z).

5.2 Motion segmentation of 2D translations

In the case of 2D translational-motion segmentation, a dynamic scene including two 2D translations (T_{a2D} and T_{b2D} with $T_{za} = T_{zb} = 0$) is considered. The analysis aims to segment points having T_{a2D} from a mixture of points having T_{a2D} and T_{b2D}. The true fundamental matrix for motion T_{a2D} is given as (from equation (5.3) with $T_{za} = 0$):

$$F_{Ta2D} = \frac{1}{f}\begin{bmatrix} 0 & 0 & T_{ya} \\ 0 & 0 & -T_{xa} \\ -T_{ya} & T_{xa} & 0 \end{bmatrix}. \tag{5.4}$$

The Sampson distances of all points (associated with translations T_{a2D} and T_{b2D}) are calculated using equation (3.4) with the substitution of F_{Ta2D} in (5.4):

$$d_{2Di} = \frac{T_{ya}(x_{2i} - x_{1i}) + T_{xa}(y_{1i} - y_{2i})}{\sqrt{2(T_{ya}^2 + T_{xa}^2)}}. \tag{5.5}$$

Substitution of point locations and their noise contamination, given in equations (5.1) and (5.2) into (5.5), yields:

$$d_{2Di} = \frac{T_{ya}(\underline{x}_{2i} + e_{ix}^2 - \underline{x}_{1i} - e_{ix}^1) + T_{xa}(\underline{y}_{1i} + e_{iy}^1 - \underline{y}_{2i} - e_{iy}^2)}{\sqrt{2(T_{ya}^2 + T_{xa}^2)}}. \tag{5.6}$$

For the distance associated with the points having T_{a2D} (denoted by d_{Ta2Di}), the above equation can be simplified to

$$d_{Ta2Di} = \frac{T_{ya}(e_{ix}^2 - e_{ix}^1) + T_{xa}(e_{iy}^1 - e_{iy}^2)}{\sqrt{2(T_{ya}^2 + T_{xa}^2)}} \sim N(0, \sigma_n^2), \tag{5.7}$$

because all the expression without noise terms are equal to zero, mandated by equation (2.3) and using the true F_{Ta2D} for the *inliers* to compute the Sampson distances.

The distances associated with the points having T_{a2D} (d_{Ta2Di}) or *inliers* in equation (5.7) are a linear combination of i.i.d. Gaussian noise variables. Therefore, they are also normally distributed with zero mean and variance σ_n^2, as the numerator and denominator cancel each other in the variance calculation.

In a similar fashion, the Sampson distances for points having T_{b2D} (denoted by d_{Tb2Di}) are derived from equation (5.6) by replacing the coordinate of the image point having T_{b2D} (5.1) and their noise components in (5.2)

$$d_{Tb2Di} = \frac{f(T_{ya}T_{xb} - T_{xa}T_{yb})}{Z_{bi}\sqrt{2(T_{ya}^2 + T_{xa}^2)}} + e, \tag{5.8}$$

where $e \sim N(0, \sigma_n^2)$ (Note that, in the above equation, the subscript b is added to the term Z_i (Z_{bi}) to indicate that it is only associated with the depth of *outliers*, i.e. the points having T_{b2D}). It can be observed from equations (5.7) and (5.8) that, the Sampson distances of the points having 2D translations T_{a2D} and T_{b2D} (d_{Ta2Di} and d_{Tb2Di}) are two Gaussian populations with the same variance but different means:

$$\begin{aligned} &N(\mu_a, \sigma_n^2) \quad \text{with} \quad \mu_a = 0 \quad \text{for } T_{a2D} \text{ and,} \\ &N(\mu_b, \sigma_n^2) \quad \text{with} \quad \mu_b = W_{2D} \quad \text{for } T_{b2D} \text{ where:} \\ &W_{2D} = \frac{f(T_{ya}T_{xb} - T_{xa}T_{yb})}{Z_{bi}\sqrt{2(T_{ya}^2 + T_{xa}^2)}\sigma_n}. \end{aligned} \tag{5.9}$$

The term W_{2D} in (5.9) represents the degree of separation between the manifolds of two translations in the parameter space. The magnitude of W_{2D} is small for similar translations and is large for different translations. Thus, W_{2D} can be used as a quantifiable measure for the similarity between two 2D translational-motions.

Alternatively, the term W_{2D} could be expressed in terms of the direction of trans-

lations T_{a2D} and T_{b2D} as

$$W_{2D} = \frac{f}{Z_{bi}\sqrt{2}\sigma_n}\|T_{b2D}\| \sin(\phi_a - \phi_b), \tag{5.10}$$

where ϕ_a and ϕ_b denote the directions of T_{a2D} and T_{b2D}, respectively (i.e. $T_{ya} = \|T_{a2D}\| \sin\phi_a$ and $T_{yb} = \|T_{b2D}\| \sin\phi_b$), and $\|T_{b2D}\|$ is the magnitude of T_{b2D} or $\|T_{b2D}\| = \sqrt{T_{yb}^2 + T_{xb}^2}$.

Generally, two Gaussian populations $N(\mu_a, \sigma_n)$ and $N(\mu_b, \sigma_n)$ would have negligible overlap when their means are more than $5\sigma_n$ away from each other. More precisely, if $|\mu_b - \mu_a| \geq 5\sigma_n$, only 0.6% of the two populations would overlap. The above amount of separation is commonly used as the threshold of correct segmentation [76, 77]. Therefore, using this threshold we assume the points having T_{a2D} will be correctly segmented if:

$$|W_{2D}| \geq 5 \quad \text{or} \quad \frac{f(T_{ya}T_{xb} - T_{xa}T_{yb})}{\bar{Z}_b\sqrt{2(T_{ya}^2 + T_{xa}^2)}\sigma_n} \geq 5. \tag{5.11}$$

In the above equation, we assume that the term Z_{bi} (the depth of points associated with object having T_{b2D} (*outliers*)) can be replaced by $\bar{Z}_b$, the average distance between the camera and the *outliers*. The justifications are twofold. First, in the target applications, the distance between the camera and the object in motion is roughly known. For example, in a traffic surveillance application, we usually know the distance between the camera and the road. Secondly, our experimental results, presented in the next section, show that the depth of the object having T_{b2D} does not have a significant effect on the segmentation performance. It is important to note that, using the term $\bar{Z}_b$ to simplify Z_{bi} (in equation (5.11)) does not change the Sampson distances associated with the target translation (T_{a2D}) and only affect the calculation

of the unwanted motion (*outliers*) residuals . The distances (d_{Ta2Di}) associated with the points having translation T_{a2D} are independent of their locations in 3D-space, as shown in equation (5.7).

In summary, we proposed the theoretical condition for segmentation (in (5.11)) of 2D translational motion. The condition is based on obtainable motion and scene parameters, in equation (5.10) i.e. the difference between translational directions (ϕ_a and ϕ_b), the level of noise (σ_n) and the desired sensitivity of the system in terms of the amount of translation. If the condition in (5.11) is satisfied, at least 99.4% of the points having T_{a2D} can be correctly segmented from points having T_{b2D}. In practice, the measurement values are always bounded and the above condition would represent perfect segmentation. In section 5.2.1 and 5.4, we will verify the validity of the condition in equation (5.11) via experiments using both synthetic and real-image data.

5.2.1 Monte Carlo experiments for 2D translational-motion segmentation

The Monte Carlo experiments for the verification of the condition for 2D translational-motion segmentation were divided into two parts. The first part of the experiments aimed to show that the term W_{2D} can be used as a measure for the degree of separation between two 2D translations. Specifically, we aimed to show that the points having T_{a2D} can be successfully segmented from the points having T_{b2D}, when the condition given in (5.11) is satisfied, i.e. $W_{2D} \geq 5$. The second part of the experiments aimed to establish sets of necessary conditions to guarantee (in terms of W_{2D}) successful

segmentation and to examine how these conditions changed when the noise level σ_n, depth of *outliers* (object having T_{b2D}) and *inlier* ratio ϵ (ratio of the number of points having target motion over the total number of points, in equation (3.2)) were varied.

In each iteration of our Monte Carlo experiments, 2000 randomly generated points in the world-coordinate system having T_{a2D} were mixed with the pairs of matching points having T_{b2D} (the number of points having T_{b2D} depends on the value of *inlier* ratio ϵ in equation (3.2)). For matching points having T_{b2D}, their X and Y coordinates were randomly generated while their Z coordinates were uniformly distributed according to $\bar{Z}_b \pm \frac{\delta_Z}{\bar{Z}_b}$ where $\frac{\delta_Z}{\bar{Z}_b} = 5\%$, 10% or 20% to represent different depth (i.e. 10%, 20% or 40%) of the object along camera optical axis. All matching points (having T_{a2D} and T_{b2D}) were projected to two images using a synthetic camera according to the camera matrix A_1 in equation (4.7), which represents a camera with a field of view around 40°, focal length of 703 pixels, principal point coordinate of (256,256) and image size of 512×512 pixels.

The points having translations T_{a2D} and T_{b2D} could be anywhere in the image plane, since the distances d_{2Di} did not depend on locations of the points (x_i and y_i) according to equations (5.7) and (5.8). All generated image-points were then perturbed with Gaussian noise of $N(0, \sigma_n)$ and the Sampson distances were calculated using equation (3.4) based on the true fundamental matrix of translation T_{a2D}. Motion segmentation was performed to identify and segment points having T_{a2D} using the MSSE [6] with d_i^2 as residuals. Although we used the segmentation step of the MSSE, the analysis was general and similar results can be expected if other robust estimators are used [42, 44].

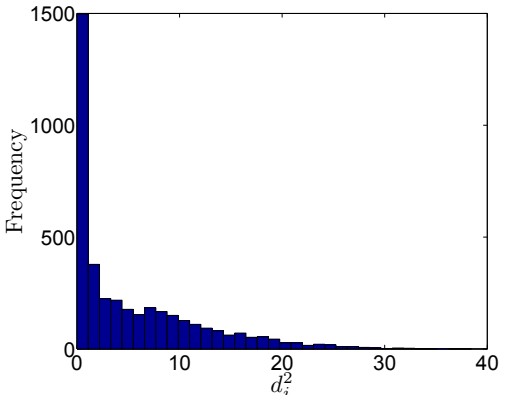
The segmentation performance was measured by the ratio (ζ) of the number of segmented points having T_{a2D} over the true number of points having T_{a2D}. The value of $\zeta = 1$ indicated correct segmentation while $\zeta > 1$ meant over-segmentation, where some of the points having T_{b2D} were segmented as points having T_{a2D}. Each experimental trial consisted of 1000 iterations and the mean and standard deviation of 1000 ζs, denoted by $\bar{\zeta}$ and σ_ζ, were recorded. These experiments were then repeated for various values of W_{2D}, *inlier* ratio ϵ, level of noise σ_n and $\frac{\delta_Z}{\bar{Z}_b}$ representing different depths of object having T_{b2D}. The pseudocode of the Monte Carlo experiments is given in figure 5.1.

Repeat ($W_{2D} = 0$ to 10) and ($\epsilon = 30\%$ to 80%).
Repeat ($\sigma_n = 0.25$ to 2) and ($\frac{\delta_Z}{\bar{Z}_b}$ from 5% to 20%).
i. Repeat ($j = 1$ to 1000).
1. Generate random T_{xa}, T_{ya}, T_{xb} and T_{yb} according to W_{2D} in (5.11).
2. Generate $N_i = 2000$ random pairs of points having T_{a2D}.
3. Generate N_o (based on (3.2)) random pairs of points having T_{b2D} with uniformly distributed Z coordinates according to $\bar{Z}_b \pm \frac{\delta_Z}{\bar{Z}_b}$.
4. Project all points on two 512×512 images using a camera with $f = 703$ pixel and $[P_x\ P_y] = [256\ 256]$.
5. Perturb all image points with Gaussian noise $N(0, \sigma_n^2)$.
6. Calculate the true F_{Ta2D} of the points having T_{a2D}.
7. Calculate the Sampson distances d_i of all points using the true F_{Ta2D}.
8. Perform segmentation using MSSE with d_i^2 as the residuals.
9. Record the segmentation performance ζ (ratio of the segmented over the true number of points having T_{a2D}).
ii. End.
iii. Calculate and record the mean and standard deviation of 1000 ζs.
End, End.

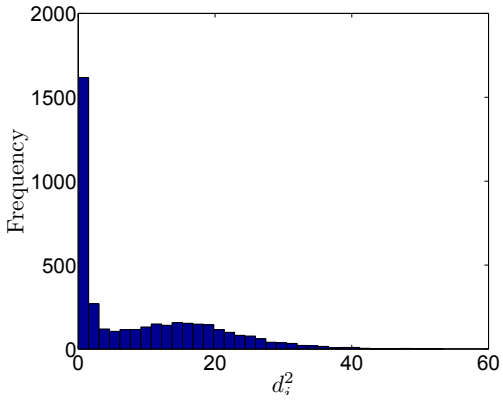
Figure 5.1: Pseudocode of the Monte Carlo experiments for the analysis of 2D translational-motion segmentation.

To demonstrate that the term W_{2D} can be used as a measure for the degree of separation, the first part of the experiments was performed by analysing the distribution of the residuals in three cases where W_{2D} is less than, close to or larger than the theoretical threshold of $W_{2D} = 5$ (in equation 5.11). The parameters for translations T_{a2D} and T_{b2D} were randomly selected such that $W_{2D} = 3$ ($W_{2D} < 5$), 4 (W_{2D} close to 5) and 6 ($W_{2D} > 5$) while the *inlier* ratio ϵ was set to 50%, the measurement noise $\sigma_n = 1$ and using the object having T_{b2D} with depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$). The histogram of the residuals associated with all image points, in terms of d_i^2s, and the segmentation performance ζ (ratio of the number of segmented points having T_{a2D} over the true points having T_{a2D}) for all cases ware recorded. Figure 5.2 shows that when W_{2D} is less than and close to five ($W_{2D} = 3$ and $W_{2D} = 4$), the values of ζs are around two, indicating incorrect segmentation. It can be observed from figure 5.2 that the segmentation is incorrect because the distributions of the residuals of the points having T_{a2D} and T_{b2D} overlap and can't be distinguished from each other. When the value of W_{2D} is larger than five ($W_{2D} = 6$), the distributions of the residuals are distinct and the value of ζ is very close to one, indicating correct segmentation, as shown in figure 5.2(c). These results were consistent with our earlier theoretical condition for segmentation in (5.11) and showed that the term W_{2D} can be used as a measure for the degree of separation between two 2D translations.

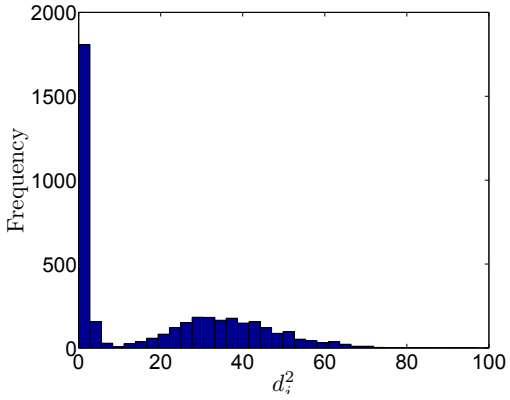
In the second part of the experiments, we examined the effect of varying parameters including W_{2D} (from 0 to 10), *inlier* ratio (from ϵ= 30% to 80%), level of noise (from σ_n= 0.25 to 1) and depth of object having T_{b2D} (from 10% to 40% or $\frac{\delta_Z}{Z_b}$= 5% to 20%) (note that the depth of object having the target translation, T_{a2D}, was random since their distances are independent to object size, depth and location as



(a) $W_{2D} < 5$ and $\zeta = 2.00$



(b) W_{2D} close to 5 and $\zeta = 2.00$



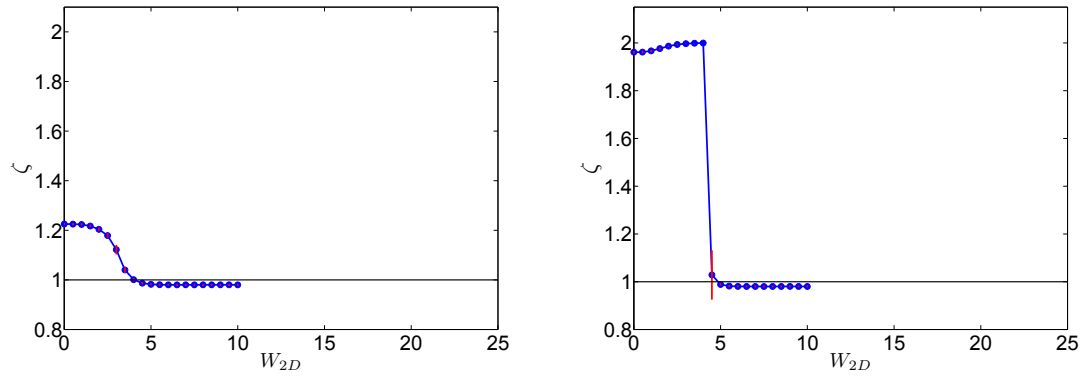
(c) $W_{2D} > 5$ and $\zeta = 0.99$

Figure 5.2: Histogram of the residuals for 2D translations when the *inlier* ratio $\epsilon = 50\%$, the measurement noise $\sigma_n = 1$ and using the object having T_{b2D} with depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$).

shown in equation (5.7)). Each experiment was repeated 1000 times and the mean $\bar{\zeta}$ and standard deviation σ_ζ of 1000 experiments were recorded. Figure 5.3 shows $\bar{\zeta}$ and σ_ζ versus W_{2D} for *inlier* ratios of 80% and 50%, while the scale of measurement noise $\sigma_n = 1$ and the depth of object having T_{b2D} is 20% ($\frac{\delta_Z}{Z_b} = 10\%$).

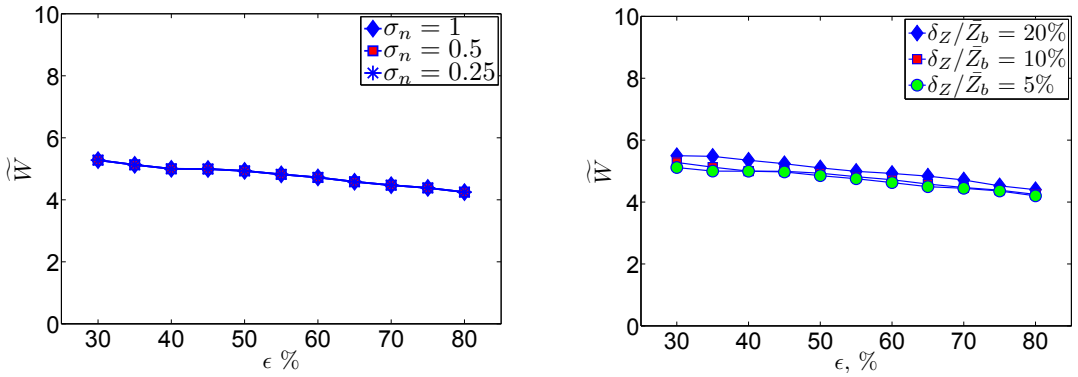
It can be observed from figure 5.3, that for small values of W_{2D}, the values of ζs are greater than one, indicating unsuccessful segmentation — some points having T_{b2D} are segmented as points having T_{a2D}. Thus, in these situations, an inaccurate *inlier-outlier* dichotomy resulted in incorrect motion-estimation and segmentation. Figure 5.3 also shows that when W_{2D} increases, the value of $\bar{\zeta}$ reduces to around 0.99 and σ_ζ reduces to zero when W_{2D} is greater than five. The value of $\bar{\zeta} \approx 0.99$ indicates that the points having T_{a2D} are correctly segmented and $\sigma_\zeta \approx 0$ confirms that the results are consistent throughout the experiments. In addition, the segmentation results are independent of the locations of points having either of the translations, since the points (X_i and Y_i) were randomly chosen in those experiments. Assuming that the points having T_{a2D} are correctly and consistently segmented when the values of $\bar{\zeta} \approx 1$ and $\sigma_\zeta \leq 0.01$, the minimum value of W_{2D} required for correct and consistent segmentation (denoted by $\widetilde{W}$) can be interpolated from figure 5.3(b) to be around $\widetilde{W} \approx 5$ when $\epsilon = 50\%$.

A broad picture of the required $\widetilde{W}$ to guarantee correct segmentation of T_{a2D} for different values of *inlier* ratio ϵ, measurement noise σ_n and depth of object having T_{b2D} $\frac{\delta_Z}{\bar{Z}_b}$ is shown in figure 5.4. Importantly, it can be observed in figure 5.4 that the values of $\widetilde{W}$ are around five, when ϵ, σ_n and $\frac{\delta_Z}{\bar{Z}_b}$ are varied in a fairly broad range. This observation indicates that *inlier* ratio, measurement noise and depth of object having T_{b2D} (*outliers*) have little effect on the segmentation performance. In addition, this observation also support the validity of using the term $\bar{Z}_b$ (average distance between camera and object having T_{b2D}) to represent the depth Z_{bi} of points associated with the object having T_{b2D} in the condition for segmentation (equation (5.11)).



(a) $\epsilon = 80\%$ (b) $\epsilon = 50\%$

Figure 5.3: Segmentation performance for 2D translational-motion segmentation from Monte Carlo experiments, when using the object having T_{b2D} with depth of 20% ($\frac{\delta_Z}{\bar{Z}_b} = 10\%$).



(a) Vary σ_n (b) Vary $\frac{\delta_z}{\bar{Z}_b}$

Figure 5.4: Conditions for segmentation of 2D translations for different values of measurement noise σ_n and depth of object having T_{b2D} ($\frac{\delta_Z}{\bar{Z}_b}$).

In summary, the results of our experiments demonstrate that the term W_{2D} in equations (5.10) and (5.11) can be used as a quantitative measure for the degree of separation between two 2D translations. They also show that, consistent with the earlier theoretical analysis, a 2D translation is guaranteed to be correctly segmented if W_{2D} is around five (5.11), irrespective of the *inlier* ratio, the measurement noise and the size and location of the translating objects.

5.3 Motion segmentation of 3D translations

In a more general case of motion segmentation involving arbitrary 3D translations including components along camera optical axis (T_a and T_b with $T_{za} \neq 0$ and $T_{zb} \neq 0$), we would again like to segment points having T_a from a mixture of points having T_a and T_b. Similar to the previous analysis, the points associated with T_a were considered as *inliers* aimed to be segmented from points having T_b (*outliers*) and the residual for the segmentation was in terms of Sampson distance measure.

The residuals of the segmentation in terms of Sampson distances d_i for all image points (associated with either T_a or T_b) can be computed using equation (3.4) with the assumption that an estimator provides the true fundamental matrix given in (5.3). Substitution of equation (5.3) into (3.4) yields

$$\begin{aligned}
d_i =& \frac{d_{2Di}(1+\alpha)}{\sqrt{1+\beta}}, \quad \text{where:} \\
\alpha =& \frac{T_{za}[y_{2i}(x_{1i}-P_x) - x_{2i}(y_{1i}-P_y) + y_{1i}P_x - x_{1i}P_y]}{f[T_{ya}(x_{2i}-x_{1i}) - T_{xa}(y_{2i}-y_{1i})]}, \\
\beta =& \frac{T_{za}^2[(x_{1i}-P_x)^2 + (y_{1i}-P_y)^2 + (x_{2i}-P_x)^2 + (y_{2i}-P_y)^2]}{2f^2(T_{ya}^2+T_{xa}^2)} \cdots \\
& - \frac{T_{za}[T_{xa}((x_{1i}-P_x)^2 + (x_{2i}-P_x)^2) + T_{ya}((y_{1i}-P_y)^2 + (y_{2i}-P_y)^2)]}{f(T_{ya}^2+T_{xa}^2)}.
\end{aligned} \tag{5.12}$$

Theoretically, a noise-free point having a particular motion should have zero Sampson distances if the true fundamental matrix of that motion is used — as indicated by equations (2.3) and (3.4), respectively. Taking the measurement noise into account, we expect that, similar to the case of 2D translations, the d_i of the points having T_a (*inliers*) will have zero mean and the same standard deviation as the measurement noise σ_n. In addition, our experimental results, presented in the following section, show that the distances of points having T_a are similar to the distribution of measurement noise regardless of translational parameters, object-size and location on the image plane. Thus, the feasibility of segmentation of the points having T_a (*inliers*) depends on the distances associated with the points having T_b (*outliers*).

For points having T_b (*outliers*), equations (5.1), (5.2), and (5.12) show that the value of d_i depends on various factors including the translations parameters (T_x, T_y, T_z for both T_a and T_b), locations of the translated points having T_b in the image plane (x_1, y_1, x_2 and y_2), measurement noise σ_n, terms associated with the distances for 2D translation (d_{2Di}), and camera parameters (f and $[P_x\ P_y]$). Derivation of a closed-form solution for the segmentation feasibility of 3D translations appears to be intractable due to the complex nature of equation (5.12). However, we can generalise and extend the condition derived for segmentation of 2D translations to the case involving multiple 3D translations by examining all parameters that affect the value of d_i associated with the point having T_b, in (5.12), and determining the ones which significantly affect the segmentation performance of points having T_a. We use the theoretical degree of separation for 2D translations (W_{2D} in equations (5.10) and (5.11)) to establish the conditions for arbitrary 3D translational-motion segmentation. These form the basis of the Monte Carlo experiments presented in section 5.3.1.

5.3.1 Monte Carlo experiments for 3D translational-motion segmentation

The Monte Carlo experiments for 3D translational-motion segmentation consisted of two main parts. The first part aimed to verify that the distribution of Sampson distances of the points having 3D translation $T_a = [T_{xa}\ T_{ya}\ T_{za}]^\top$ (the *inliers* or target translation) is similar to the distribution of the measurement noise. The second part of the experiments was designed to evaluate the usefulness of the term W_{2D} as the degree of separation between two 3D translations, and to develop sets of sufficient conditions that guarantee successful segmentation of a 3D translation. In addition, we examined the effect of the following quantities, as shown in (5.12), on successful segmentation:

- the magnitude and the direction of T_z (for both translations T_a and T_b), the scale of noise and the distance between camera and object associated with *outliers* T_b (merged in one quantity $\frac{T_z}{\bar{Z}_b \sigma_n}$),
- the size, depth and location of object/points having translation T_b,
- the camera parameters and
- the *inlier* ratio ϵ.

The 3D Monte Carlo experimental set-up was similar to the 2D case and started by mixing 2000 randomly selected points in the world-coordinate system having 3D translation T_a, with the pairs of matching points having another 3D translation (i.e. T_b). For points having T_b, the Z coordinates were uniformly distributed according to $\bar{Z}_b \pm \frac{\delta_Z}{\bar{Z}_b}$ to represent different object depth or the size of object in Z direction. Then,

all matching points (having T_a or T_b) were projected to two images using a synthetic camera according to A_1 in (4.7). Throughout the experiment, the camera parameters were changed by varying the focal length f and image principal point $[P_x\ P_y]$ (with a smaller focal length, the camera has a higher focus for the same $\bar{Z}_b$ and the objects appear larger on the image, thus increasing the effect of motion in the image). We assumed that the images were square with the principal point $[P_x\ P_y]$ located at their center, hence larger values of $[P_x\ P_y]$ represented images with larger dimension.

In many computer-vision applications, image points having a motion are associated with a moving object and are largely confined to one part of an image. As such, to examine the effect of the size and location of object having T_b (*outliers*), we designed our experiments in such a way that the points having T_b were confined to a square. The length of each side of the square was denoted by $l\%$ of the image size and the symbol D_p denoted the distance from the center of the square to the image principal point. Higher values of l (size) and $\frac{\delta_Z}{\bar{Z}_b}$ (depth) represented a larger 3D object as its associated feature points cover a larger area in the image while larger magnitude of D_p indicated that the points were located further away from image principal point.

The ground-truth for the matching pairs were perturbed with Gaussian noise of $N(0, \sigma_n)$, the Sampson distances were calculated using equation (3.4) based on the true F_{Ta}, and segmentation was performed using the MSSE [6]. To quantify the segmentation performance, we computed the ratio of the number of segmented points having T_a over the true number of points having T_a (denoted by ζ). Each experiment was repeated 1000 times and the mean ($\bar{\zeta}$) and the standard deviation (σ_ζ) of segmentation-performance measure were recorded. The magnitude and direction of T_z (in both T_a and T_b), the scale of measurement noise and the distance between

camera and object associated with *outliers* T_b (merged into one parameter $\frac{T_z}{Z_b \sigma_n}$), the *inlier* ratio ϵ, the size, depth and location of object having T_b (parameterised by l, $\frac{\delta_Z}{Z_b}$ and D_p), and the camera parameters (f and $[P_x \ P_y]$) were all varied to measure their effect on the segmentation performance. The pseudocode for the Monte Carlo experiments is shown in figure 5.5.

In the first part of the experiments, we recorded the histogram of d_i associated with random points having T_a and compared it with the distribution of measurement noise. Figure 5.6 shows the histograms of d_i for randomly selected points (for all X, Y and Z coordinate in 3D-space) having a random translation T_a when the measurement noise is distributed according to a normal distribution with $\sigma_n = 1$, while the object size l and location D_p on the images are varied (l from 20% to 40% of the image and D_p from 10% to 30% of the image). We also recorded the mean and standard deviation of d_i associated with the points having T_a, denoted by the symbols M and S, respectively. It can be observed from figure 5.6 that the histograms of d_i for the the points having T_a appear as normal distributions and the values of M and S for both cases are very close to the mean and standard deviation of measurement noise (zero mean and $\sigma_n = 1$, respectively).

The above experiments were repeated 1000 times for different size (l) of object having T_a, measurement noise signal ($N(0, \sigma_n)$) and random location D_p. We calculated the statistical mean and standard deviation of the 1000 recorded values of M and S (denoted by $\bar{M}$, σ_M, $\bar{S}$ and σ_S) as shown in table 5.1. The values of $\bar{M}$ and $\bar{S}$ for the points having T_a are overall very close to the mean and standard deviation of the measurement noise (zero mean and $\bar{S} \approx \sigma_n$). In addition, the standard deviation of M and S values (σ_M and σ_S) are very close to zero, indicating consistent values

Repeat ($W_{2D} = 0$ to 120) and ($\epsilon = 30\%$ to 80%).
Repeat ($\frac{T_z}{\bar{Z}_b \sigma_n} = 5\%$ to 30%) and (directions of T_{za} and T_{zb}).
Repeat ($l = 20\%$ to 40%), ($D_p = 10\%$ to 30%) and ($\frac{\delta_Z}{\bar{Z}_b} = 5\%$ to 20%).
Repeat (f = 500, 703 and 900 pixel), ($[P_x \ P_y] = [200\ 200]$, [256 256] and [300 300]) and (image size 400×400, 512×512 and 600×600).

i. Repeat ($j = 1$ to 1000).
   1. Generate random T_{xa}, T_{ya}, T_{xb} and T_{yb} according to W_{2D} in (5.11).
   2. Generate $N_i = 2000$ random pairs of points having T_a with $T_{za} = \frac{T_z}{\bar{Z}_b \sigma_n}$.
   3. Generate N_o (3.2) random pairs of points (with uniformly distributed Z coordinates according to $\bar{Z}_b \pm \frac{\delta_Z}{\bar{Z}_b}$) having T_b with $T_{zb} = \frac{T_z}{\bar{Z}_b \sigma_n}$.
   4. Project all points on two images using a camera with focal length f and principal point $[P_x \ P_y]$.
   5. Crop the points having T_b so that they are within $l\% \times l\%$ of the image and located at $D_p\%$ of the image.
   6. Perturb all image points with Gaussian noise $N(0, \sigma_n^2)$.
   7. Calculate the true F_{Ta} of the points having T_a.
   8. Calculate the Sampson distances d_i of all points using the true F_{Ta}.
   9. Perform segmentation using MSSE with d_i^2 as the residuals.
   10. Record the segmentation performance ζ (ratio of the segmented over the true number of points having T_a).

ii. End.
iii. Calculate and record the mean and standard deviation of 1000 ζs.

End, End, End, End.

Figure 5.5: Pseudocode of the Monte Carlo experiments for 3D translational-motion segmentation.

of M and S throughout the experiments. The above results demonstrate that, irrespective of the size, depth and location of an object in an image, the distribution of residuals for the points having T_a is identical to the distribution of the measurement noise when the latter is Gaussian.

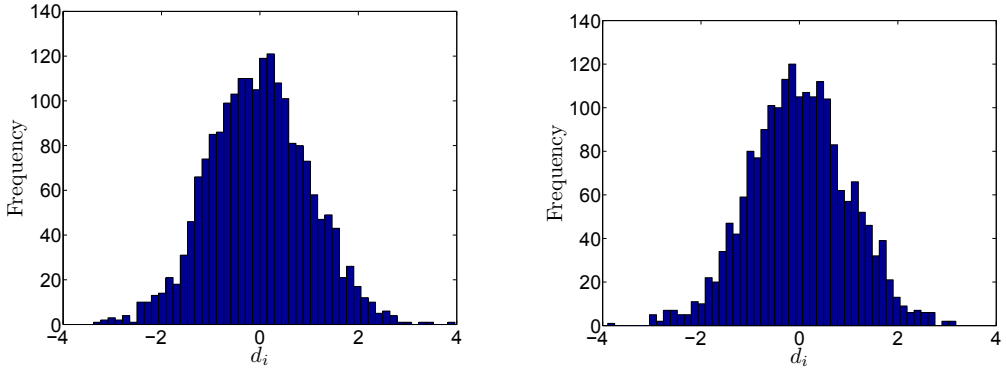
The second part of the Monte Carlo experiments focused on evaluating the usefulness of W_{2D} as a measure for the degree of separation between two 3D translations and developing a set of sufficient conditions for successful segmentation of a transla-

tion (T_a) from another translation (T_b). In this part, we examined the effect of the following factors on the validity of the proposed conditions:

- the magnitude and direction of T_z (for both T_a and T_b),
- the scale of measurement noise and the distance between camera and object associated with *outliers* T_b (merged into $\frac{T_z}{Z_b \sigma_n}$),
- the *inlier* ratio ϵ,
- the camera parameters (f and $[P_x \ P_y]$), and
- the size, depth and location of the object having T_b (l, $\frac{\delta_Z}{Z_b}$ and D_p) (note that the size, depth and location of points associated with the object having target translation, T_a, was randomly selected since their distances are according to zero mean and standard deviation of σ_n as shown in the earlier results).

The segmentation performance was again measured in terms of ζ (ratio of the number of segmented over the true number of points having T_a) and its mean and standard deviation ($\bar{\zeta}$ and σ_ζ) were also recorded for all iterations.

Figure 5.7 shows the segmentation performance ζ versus W_{2D} when the parameter $\frac{T_z}{Z_b \sigma_n} = -5\%$ and using the object having T_b with size ($l = 30\%$), depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$), located at $D_p = 10\%$ while the scene is viewed by a camera with $f = 703$ and image principal point $[P_x \ P_y] = [256 \ 256]$. In both cases, it can be observed that, for small W_{2D} values, some of the points associated with T_b are incorrectly segmented as having T_a indicated by $\zeta > 1$. By increasing W_{2D}, the segmentation performance indicators $\bar{\zeta}$ and σ_ζ are converged and reduced to around 0.99 and zero, respectively.



(a) $M = -0.02$, $S = 1.01$ (b) $M = -0.01$, $S = 1.00$

Figure 5.6: Histogram of residuals associated with points having random T_a when $\sigma_n = 1$ and located at $D_p = 20\%$ with size $l = 20\%$ in (a) and $D_p = 10\%$ and $l = 30\%$ in (b).

This demonstrates the usefulness of the term W_{2D} as a measure for the degree of separation for 3D translational-motion segmentation.

To develop the conditions to guarantee correct segmentation of 3D translational motions, in terms of the required W_{2D}, we assumed that a correct and consistent segmentation was prescribed by $\bar{\zeta} \approx 1$ and $\sigma_\zeta \leq 0.01$. From figure 5.7(a), the minimum required value of W_{2D} for correct segmentation, denoted by $\widetilde{W}$, is around 11 if the value of the *inlier* ratio is around 80%.

Broad pictures of the $\widetilde{W}$ value required to guarantee correct segmentation of the points having T_a with different motion parameters ($\frac{T_z}{Z_b \sigma_n}$ and direction of T_z for both T_a and T_b), size, depth and location (l, $\frac{\delta_Z}{Z_b}$ and D_p) of object having T_b, camera parameters (f and $[P_x \; P_y]$) and *inlier* ratio ϵ are shown in figures 5.8, 5.9 and 5.10. The results in figure 5.8 show that the segmentation becomes more challenging and

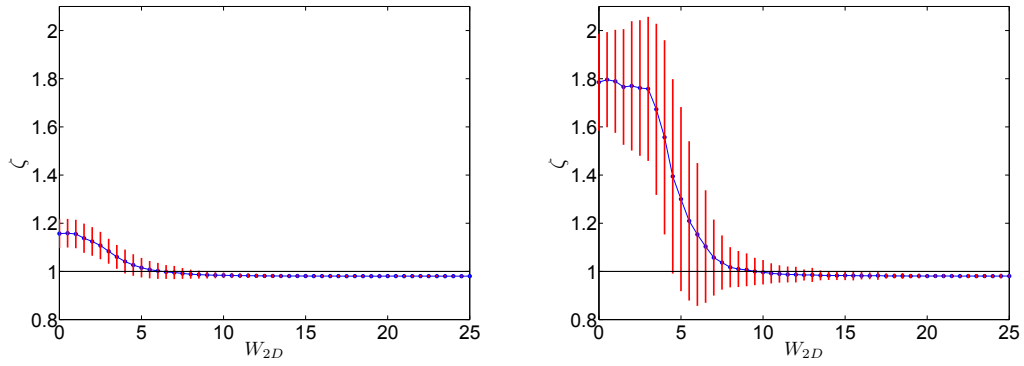
Table 5.1: Mean and standard deviation of M and S associated with random points having random translation T_a and location, when the size of object having T_a (l) and measurement noise (σ_n) were varied.

σ_n	l	20%	30%	40%	50%	Random
1	$\bar{M}$	-0.007	-0.006	-0.005	-0.004	-0.001
	σ_M	0.016	0.022	0.025	0.016	0.022
	$\bar{S}$	0.997	0.999	0.991	1.004	1.000
	σ_S	0.015	0.014	0.013	0.011	0.016
0.5	$\bar{M}$	0.000	0.001	0.000	0.000	0.000
	σ_M	0.011	0.011	0.011	0.010	0.011
	$\bar{S}$	0.500	0.500	0.500	0.500	0.500
	σ_S	0.008	0.008	0.009	0.008	0.008
0.25	$\bar{M}$	0.000	0.000	0.000	0.000	0.000
	σ_M	0.005	0.006	0.005	0.005	0.006
	$\bar{S}$	0.250	0.250	0.250	0.250	0.250
	σ_S	0.004	0.004	0.004	0.004	0.004

difficult, as indicated by larger values of the required $\widetilde{W}$, when there are many points having T_b (small value of ϵ), or when the location of points having T_b are farther away from the image principal point (large value of D_p), or when the effect of T_z is more pronounced (large value of $|\frac{T_z}{\bar{Z}_b\sigma_n}|$). Importantly, by comparing figure 5.8(a) to 5.9(a), 5.8(a) to 5.9(b) and 5.8(b) to 5.10, we observe that the size and depth (l and $\frac{\delta_Z}{\bar{Z}_b}$) of object having T_b (*outliers*), the camera parameters (f and $[P_x\ P_y]$) and the direction of T_z (in both T_a and T_b) have little effect on the segmentation performance of points

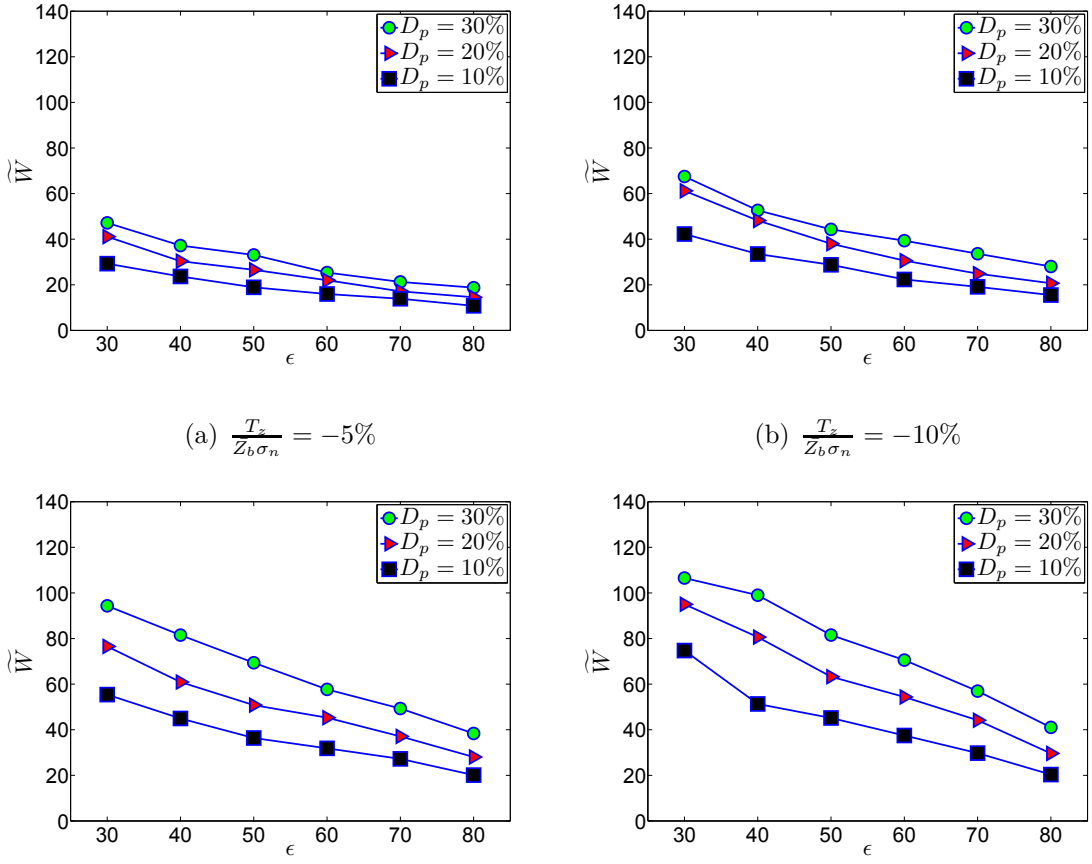
having translation T_a.

To provide an insight into the predictive capability of the term $\widetilde{W}$ in a particular situation, we analysed the distribution of residuals in three cases where W_{2D} is less than, close to or larger than the required threshold. The selected threshold is $\widetilde{W} \approx 19$ when $\frac{T_z}{Z_b \sigma_n} = -5\%$, $\epsilon = 50\%$, size of object having T_b (*outliers*) $l = 30\%$ and $\frac{\delta_Z}{Z_b} = 10\%$ located at $D_p = 10\%$ of the image, extracted from figure 5.8(a). Motion segmentations were performed to recover points having T_a when the parameters of translations T_a and T_b were randomly selected in such a way that $W_{2D} = 5$ ($W_{2D} < \widetilde{W}$), $W_{2D} = 17$ ($W_{2D} \approx \widetilde{W}$) and $W_{2D} = 22$ ($W_{2D} > \widetilde{W}$) while $\frac{T_z}{Z_b \sigma_n} = -5\%$, $\epsilon = 50\%$, $l = 30\%$, $\frac{\delta_Z}{Z_b} = 10\%$ and $D_p = 10\%$. The histogram of the residuals for all image points (T_a and T_b) and their corresponding values of segmentation performance ζ for all cases



(a) $\epsilon = 80\%$ (b) $\epsilon = 50\%$

Figure 5.7: Segmentation performance for 3D translational-motion segmentation when $\frac{T_z}{Z_b \sigma_n} = -5\%$ from Monte Carlo experiments, using the object having T_b with size ($l = 30\%$), depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$) and located at $D_p = 10\%$ from image principal point.



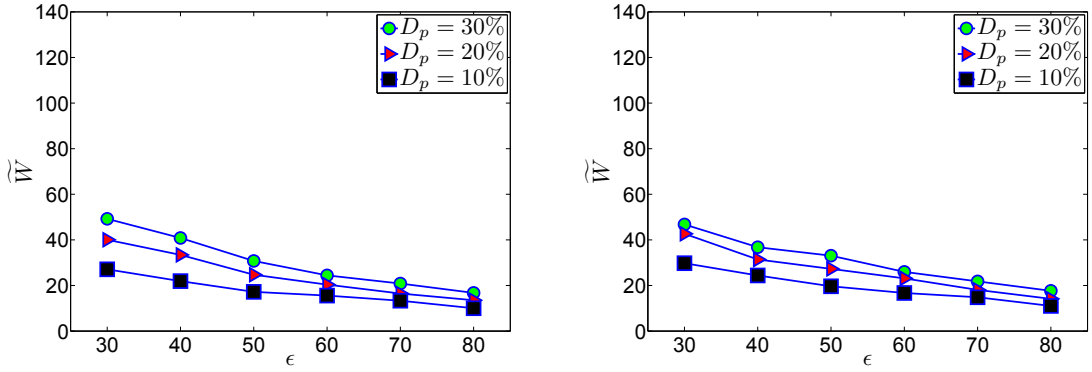
(a) $\frac{T_z}{Z_b \sigma_n} = -5\%$

(b) $\frac{T_z}{Z_b \sigma_n} = -10\%$

(c) $\frac{T_z}{Z_b \sigma_n} = -20\%$

(d) $\frac{T_z}{Z_b \sigma_n} = -30\%$

Figure 5.8: Conditions for 3D translational-motion segmentation for various $\frac{T_z}{Z_b \sigma_n}$ using the object having T_b with size ($l = 30\%$), depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$) and camera parameters of $f = 703$ and $[P_x \ P_y] = [256 \ 256]$.



(a) $f = 900$ and $(P_x\ P_y) = (300\ 300)$ (b) $l = 40\%$ and $\frac{\delta_Z}{Z_b} = 20\%$

Figure 5.9: Conditions for 3D translational-motion segmentation for $\frac{T_z}{Z_b \sigma_n} = -5\%$ when camera parameters and size, depth and location of object having T_b are varied.

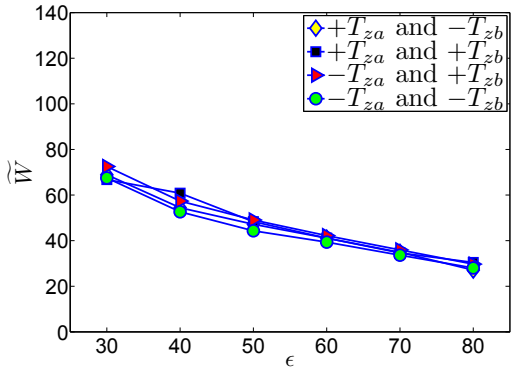
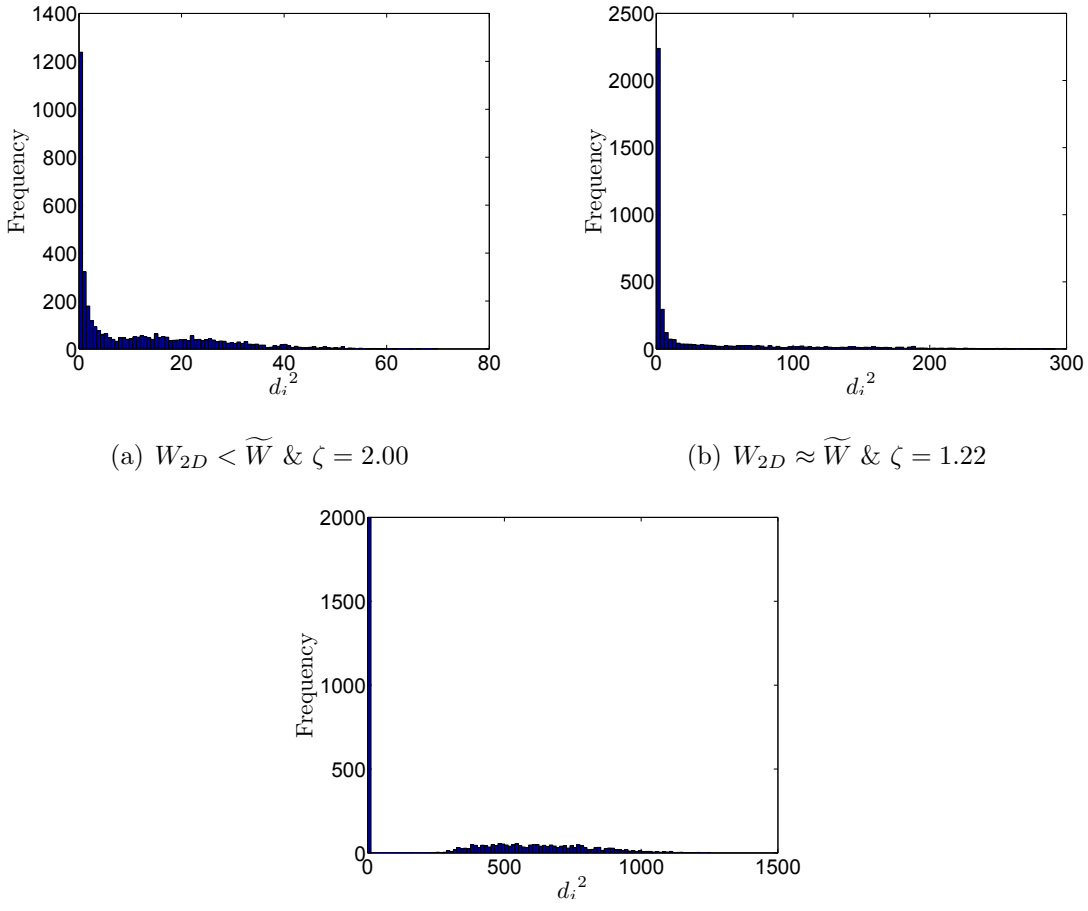


Figure 5.10: Conditions for 3D translational-motion segmentation for all directions of T_z when $\frac{T_z}{Z_b \sigma_n} = 10\%$ using the object having T_b with size ($l = 30\%$), depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$), located at $D_p = 30\%$ from image principal point and camera parameters of $f = 703$ and $[P_x\ P_y] = [256\ 256]$.



(a) $W_{2D} < \widetilde{W}$ & $\zeta = 2.00$

(b) $W_{2D} \approx \widetilde{W}$ & $\zeta = 1.22$

(c) $W_{2D} > \widetilde{W}$ & $\zeta = 0.98$

Figure 5.11: Histogram of the residuals for 3D translational-motion segmentation when $\frac{T_z}{Z_b \sigma_n} = -5\%$ and *inlier* ratio $\epsilon = 50\%$ using the object having T_b with size ($l = 30\%$), depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$) and located at $D_p = 10\%$ from image principal point.

are shown in figure 5.11. The results in figures 5.11(a) and 5.11(b) show that when W_{2D} is less than or close to $\widetilde{W}$, the segmentations are unsuccessful ($\zeta \geq 1$) and the distribution of the residuals associated with T_a and T_b are too close to be distinguished from each other. When the value of W_{2D} is greater than $\widetilde{W}$, the distributions of the residuals are distinct and the value of ζ is around one, which indicates correct segmentation as shown in figure 5.11(c).

These results show the term W_{2D} can be used as a measure for the degree of separation between two 3D translations. The derived conditions, in terms of the required W_{2D} ($\widetilde{W}$) for segmentation can be applied as a performance predictor for translational-motion segmentation. In addition, the proposed sufficient conditions for segmentation were not significantly affected by the variation of camera parameters, direction of T_z, size and depth of *outliers*. In practice, the term W_{2D} can be estimated using equation (5.10) based on the difference between translation angles, the scale of noise and the desired sensitivity of the system in terms of the amount of translation. Hence, we can predict the outcome of a translational-motion segmentation problem for the expected *inlier* ratio of a scene. Therefore, these conditions provide practical guidelines for practitioners in designing motion-segmentation solutions for computer-vision applications.

5.4 Experiments using real images

Experiments using real-image data were set up to demonstrate the relevance of the conditions for segmentation derived from the Monte Carlo experiments. We envisaged a scene containing two objects having either two different 2D translations (T_{a2D} and

T_{b2D}) or 3D translations (T_a and T_b). The points associated with T_{a2D} or T_a were the target translation (*inliers*) whereas the points having T_{b2D} or T_b were the unwanted translations (*outliers*).

The experimental aim is to investigate the theoretical limit of motion segmentation involving 2D and 3D translational motions. The effect of imperfect estimate of the fundamental matrix to the conditions for segmentation is beyond the scope of this work. In practice, the fundamental matrix can be accurately estimated using a number of robust methods [3, 101, 124] and the *gross outliers* can be removed by applying robust estimators as part of motion segmentation process [123]. The estimation issues including the estimation of fundamental matrix in terms of feasibility and accuracy, have already been thoroughly analysed [42, 44]. As such in our experiments using real-image data — identical to our earlier theoretical analysis and Monte Carlo experiments — we assumed that an accurate estimate of the fundamental matrix of T_{a2D} or T_a was provided by a robust estimator and there were no mismatches (*gross outliers*) in the image data. Thus, the fundamental matrix of T_{a2D} or T_a was calculated using equation (2.5) and occasional *gross outliers* were manually removed from the data. These assumptions needed to be taken in order to eliminate the effect of potential errors from the estimation of the fundamental matrix and the presence of *gross outliers*, to the experimental results.

In our experiments, a camera was calibrated using a publicly available camera-calibration toolbox [16] to determine the camera parameters in terms of focal length, the image principal point and image distortion. These parameters were used to calculate the fundamental matrix of the target motions and to determine the location of the object (D_p) — associated with the unwanted motion (*outliers*) — with respect

to the image principal point.

A specially designed and fabricated triangular-shaped 3D object was used as the target object (object-a) and another two similar objects (object-b_1 and object-b_2) as the unwanted objects. The side of each object that was visible to the camera consisted of non-repeating patterns to ensure maximum number of image point can be extracted from their images. In addition, the non-repeating patterns reduced the possibility of having *gross outliers* in the data. The dimensions of object-b_1 and object-b_2 were designed such that the value of $\frac{\delta_Z}{Z_b}$ was around 10% to represent object with depth of around 20% for object-b_1 and $\frac{\delta_Z}{Z_b} \approx 5\%$ or depth around 10% for object-b_2.

Camera images were recorded, before and after, object-a and either object-b_1 or object-b_2 were displaced according to various pairs of 2D translation (T_{a2D} and T_{b2D}) or 3D translation (T_a and T_b). The distortions of all images were reduced using radial and tangential distortion models suggested by the camera-calibration toolbox [16]. The corresponding feature points in each pair of images were extracted and determined using the Scale-Invariant Feature Transform (SIFT) algorithm [58, 56]. The *inlier* ratio ϵ in each pair of images was varied from 30% to 80% by removing some of the points having unwanted translations (T_{b2D} and T_b) while maintaining the points having the target translations (T_{a2D} and T_a). In order to vary the relative size of object having 3D translation T_b, the points associated with T_b (*outliers*) were cropped to be within the appropriate part (i.e. $l \times l \approx 30\% \times 30\%$ or $40\% \times 40\%$) of the image.

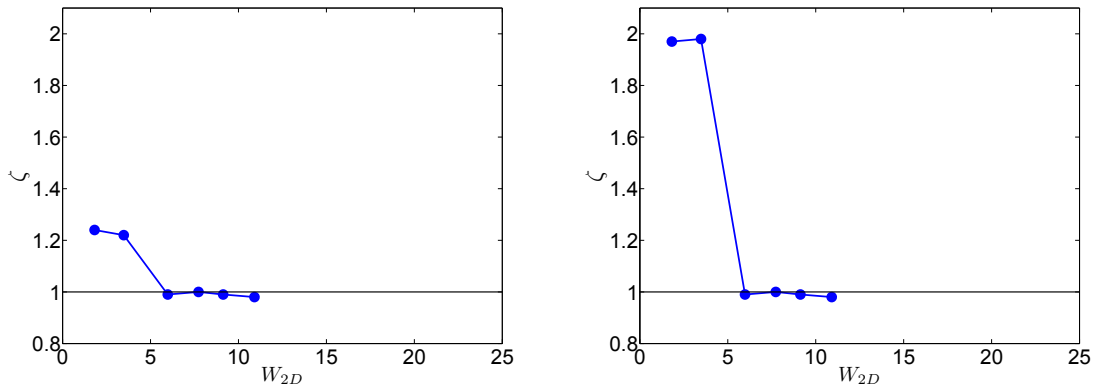
The fundamental matrix associated with target translations (T_{a2D} or T_a) and the Sampson distances for all points in each pair of images were calculated using equations (2.5) and (3.4) by substitution of known T_{a2D} or T_a and the estimated camera matrix.

Segmentation was performed to identify the points associated with T_{a2D} or T_a from the points having T_{b2D} or T_b, respectively, in each pair of images using the MSSE [6]. The ratio of the segmented points over the true number of points having T_{a2D} or T_a (ζ) was recorded and the noise level σ_n was estimated using equation (3.5) for each pair of images. The experiments involving 2D translations were repeated using the pair of object-a and object-b_2 (depth of around 10% or $\frac{\delta_Z}{Z_b} \approx 5\%$) to see the effect of changing the depth of object having T_{b2D} on the segmentation results. Meanwhile, experiments for 3D translational-motions segmentation were repeated for the different values of: $\frac{T_z}{Z_b \sigma_n}$ (for both T_a and T_b), size and location of object, having translation T_b, (l and D_p) in the image plane.

Figures 5.12 and 5.13 show the values of segmentation performance ζ versus W_{2D} for both cases involving 2D and 3D translational motions when the depth of object having T_{b2D} or T_b is around 20% ($\frac{\delta_Z}{Z_b} \approx 10\%$) and *inlier* ratio $\epsilon = 80\%$ or 50%. For the 3D case in figure 5.13, the size and location of object having T_b were selected to be around $l \approx 30\%$ and $D_p \approx 10\%$ of the image, while the estimated scale of noise $\sigma_n \approx 0.6$ pixel (estimated using equation (3.5)) and $\frac{T_z}{Z_b \sigma_n} \approx -5\%$ for both T_a and T_b.

We observed a similar trend of the segmentation performance ζ in both the results of Monte Carlo and real-image experiments as shown in figures 5.3(a) and 5.12(a), 5.3(b) and 5.12(b), 5.7(a) and 5.13(a) and 5.7(b) and 5.13(b) when $\epsilon = 80\%$ and 50%. Intuitively, the segmentation of points having T_{a2D} and T_a are incorrect ($\zeta > 1$) when the value of W_{2D} is small, and it improves ($\zeta \approx 1$) with the larger value of W_{2D}. The conditions for segmentation in terms of the minimum W_{2D} ($\widetilde{W}$) were extracted assuming that a correct segmentation was mandated by the value of $\zeta \approx 1$. From figures 5.12(b) and 5.13(b), these conditions from real image data are $\widetilde{W} \approx 6$ for 2D

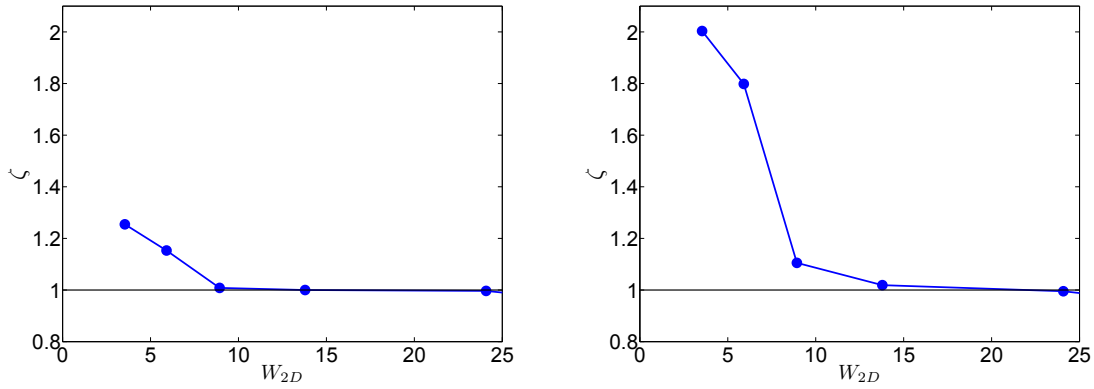
translations and $\widetilde{W} \approx 22$ for 3D translations when the *inlier* ratio is around 50%.



(a) $\epsilon = 80\%$ (b) $\epsilon = 50\%$

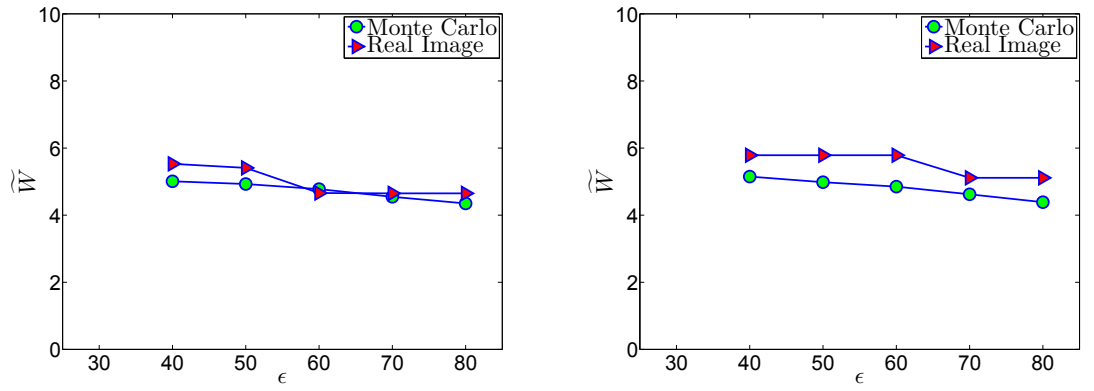
Figure 5.12: Segmentation performance for 2D translational-motion segmentation from experiments using real images, when the depth of object having T_{b2D} is around 20% $\frac{\delta_Z}{Z_b} \approx 10\%$.

The conditions for segmentation of 2D and 3D translations observed in these experiments were comparable to the relevant results from the Monte Carlo experiments and are shown in figures 5.14, 5.15 and 5.16 for the different values of $\frac{T_z}{Z_b \sigma_n}$, the size, depth and location of objects having T_{b2D} or T_b (l, $\frac{\delta_Z}{Z_b}$ and D_p). It was observed that, consistent with the previous results of the Monte Carlo experiments, the magnitude of $\widetilde{W}$ increases while ϵ decreases and does not significantly depend on the size and depth of object having T_{b2D} and T_b (l and $\frac{\delta_Z}{Z_b}$) for both 2D (figure 5.14) and 3D translations (figure 5.16). In addition, for the case of 3D translational-motion segmentation, the value of $\widetilde{W}$ also increases with increasing $|\frac{T_z}{Z_b \sigma_n}|$ or increasing D_p, as shown in figure 5.15 and from the comparison of figures 5.15(a) and 5.16(a). This means that the



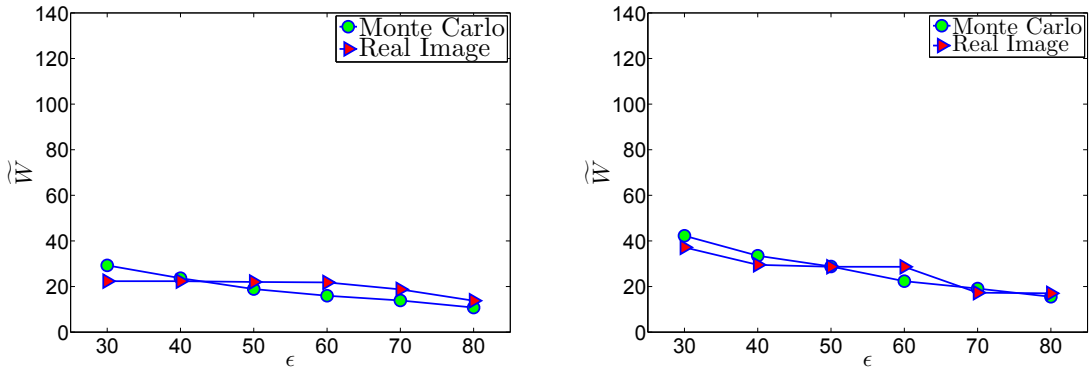
(a) $\epsilon = 80\%$ (b) $\epsilon = 50\%$

Figure 5.13: Segmentation performance for 3D translational-motion segmentation when $\frac{T_z}{Z_b \sigma_n} \approx -5\%$ from experiment using real images. The size, depth and location of object having T_b are $l \approx 30\%$, 20% or $\frac{\delta_Z}{Z_b} \approx 10\%$ and $D_p \approx 10\%$ of the image.



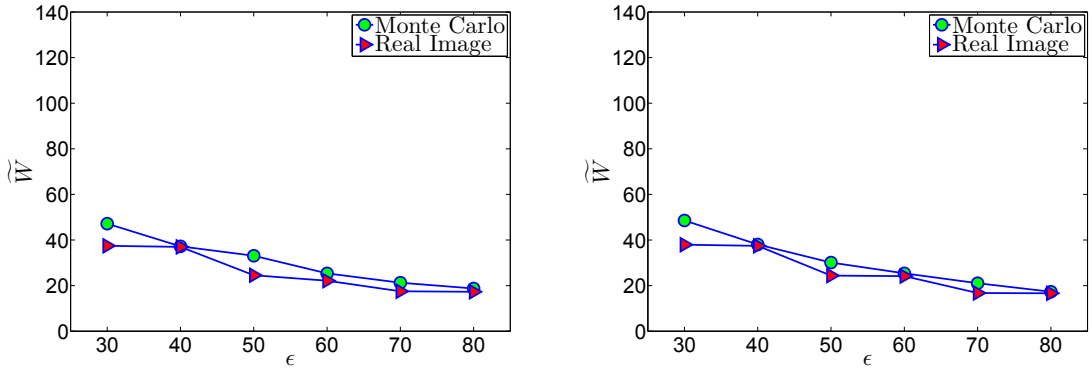
(a) depth of 10% or $\frac{\delta_Z}{Z_b} \approx 5\%$ (b) depth of 20% or $\frac{\delta_Z}{Z_b} \approx 10\%$

Figure 5.14: Conditions for 2D translational-motion segmentation from Monte Carlo experiments and real-image data for different depth $\frac{\delta_Z}{Z_b}$ of object having T_{b2D}.



(a) $\frac{T_z}{Z_b\sigma_n} \approx -5\%$ (b) $\frac{T_z}{Z_b\sigma_n} \approx -10\%$

Figure 5.15: Conditions for 3D translational-motion segmentation when $\frac{T_z}{Z_b\sigma_n} \approx -5\%$ are varied. The size, depth and location of object having T_b are according to $l = 30\%$, 20% or $\frac{\delta_Z}{Z_b} \approx 10\%$ and $D_p = 10\%$ of the image.



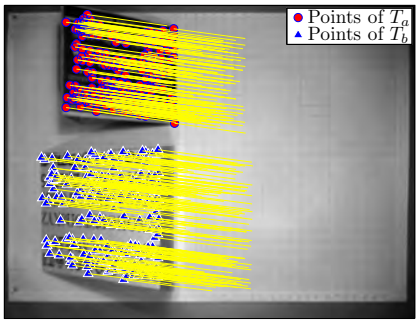
(a) $l \approx 30\%$ and $\frac{\delta_Z}{Z_b} \approx 10\%$ (b) $l \approx 40\%$ and $\frac{\delta_Z}{Z_b} \approx 10\%$

Figure 5.16: Conditions for 3D translational-motion segmentation when the object-sizes are varied. The parameter $\frac{T_z}{Z_b\sigma_n} \approx -5\%$ and object-location is according to $D_p = 30\%$ of the image.

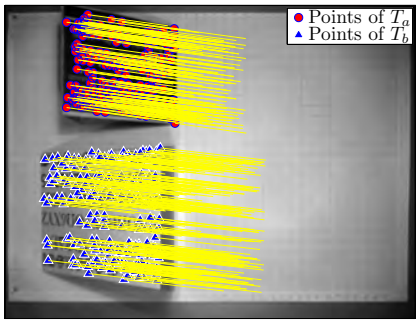
segmentation of points having T_a becomes more difficult when the value of $|\frac{T_z}{Z_b\sigma_n}|$ is large or when the points having T_b are located farther away from the image principal point.

To provide some insight into the applicability of the proposed conditions, we examined the histogram of residuals and the segmentation performance when W_{2D} was either less or greater than the required thresholds. The selected thresholds were $\widetilde{W} \approx 5$ for 2D case (when $\epsilon = 50\%$ and $\frac{\delta_Z}{Z_b} = 10\%$) and $\widetilde{W} \approx 33$ for 3D case ($\epsilon = 50\%$, $\frac{T_z}{Z_b\sigma_n} \approx -5\%$, $l = 30\%$, $\frac{\delta_Z}{Z_b} \approx 10\%$ and $D_p = 30\%$), both from Monte Carlo experimental results in figures 5.4 and 5.8(a). Figures 5.17(c) and 5.18(c) show that for W_{2D} less than $\widetilde{W}$ ($W_{2D} \approx 3.5$ for 2D translations and $W_{2D} \approx 10$ for 3D translations) the distributions of residuals are not distinguishable and the values of ζ are around two, indicating incorrect segmentations. However, when the value of W_{2D} is increased to be greater than $\widetilde{W}$ ($W_{2D} \approx 9$ for 2D translations and $W_{2D} \approx 41$ for 3D translations), the points having T_a or T_{a2D} are correctly segmented ($\zeta \approx 1$), as shown in figures 5.17(d) and 5.18(d), respectively. The segmentation is successful because the residuals associated with the target motion, i.e. T_{a2D} or T_a, can be distinguished from the overall population of d_i^2, as shown in figures 5.17(f) and 5.18(f).

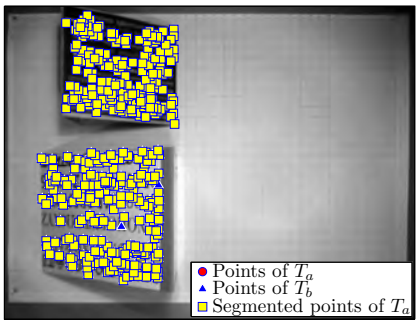
These results show the capability of the proposed conditions to guarantee successful segmentation and enable one to correctly predict the success of a segmentation scenario. They also show that the Monte Carlo experimental results are very relevant to the problem encountered in real-world applications.



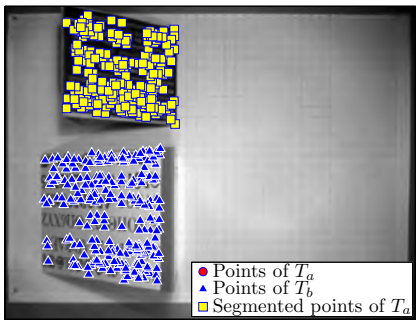
(a) $W_{2D} < \widetilde{W}$, $W_{2D} = 3.5$



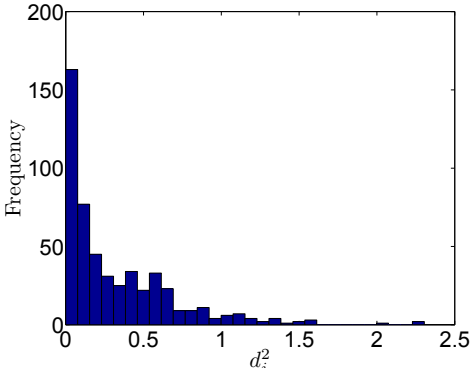
(b) $W_{2D} > \widetilde{W}$, $W_{2D} = 9.1$



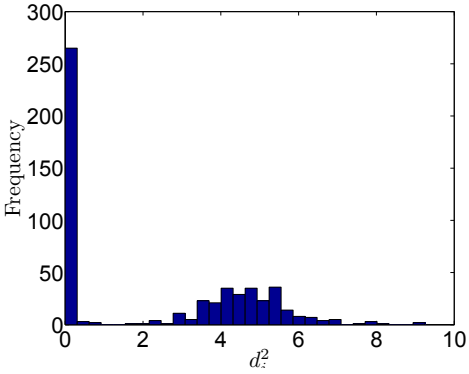
(c) $\zeta = 1.98$



(d) $\zeta = 0.99$

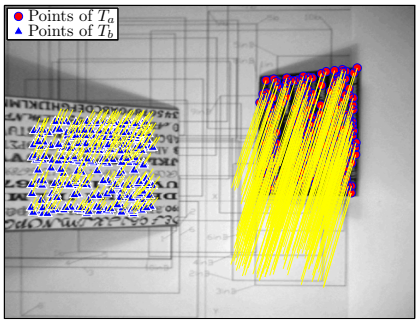


(e) $\zeta = 1.98$

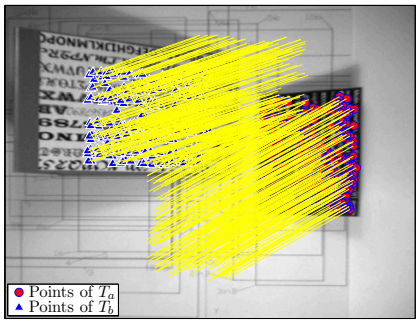


(f) $\zeta = 0.99$

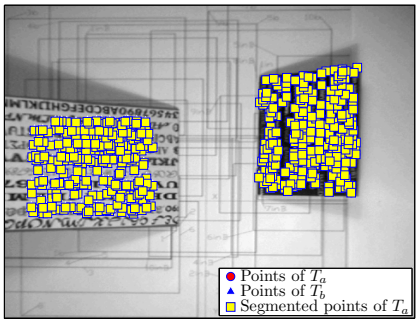
Figure 5.17: The ground-truth points having T_{a2D} and T_{b2D} are superimposed onto first image ((a) and (b)) when $\epsilon = 50\%$ and $\frac{\delta_Z}{Z_b} = 10\%$. Segmented points ((c) and (d)) and the histogram for residuals ((e) and (f)).



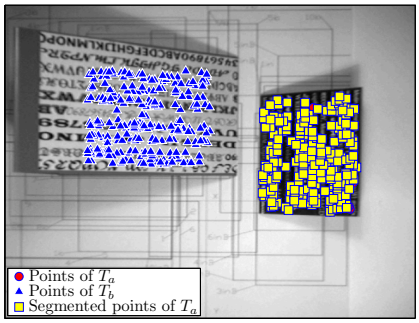
(a) $W_{2D} < \widetilde{W}$, $W_{2D} = 10.4$



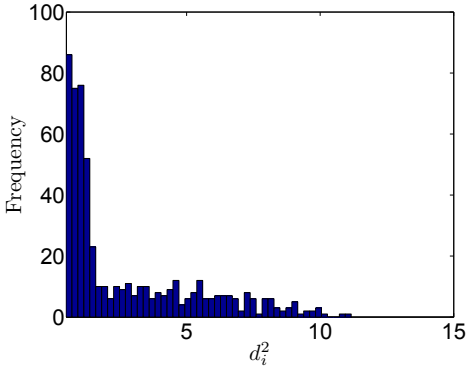
(b) $W_{2D} > \widetilde{W}$, $W_{2D} = 40.5$



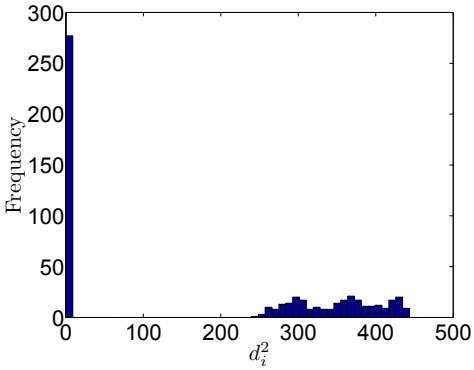
(c) $\zeta = 2.00$



(d) $\zeta = 0.98$



(e) $\zeta = 2.00$



(f) $\zeta = 0.98$

Figure 5.18: The ground-truth points having T_a and T_b are superimposed onto first image ((a) and (b)) when $\epsilon = 50\%$, $\frac{T_z}{Z_b \sigma_n} \approx -5\%$, $D_p = 30\%$, $l = 30\%$ and $\frac{\delta_Z}{Z_b} = 10\%$. Segmented points ((c) and (d)) and the histogram for residuals ((e) and (f)).

5.5 Conclusion

A measure for the degree of separation (W_{2D} in (5.10)) between two 2D translations was theoretically derived to analyse the feasibility of translational-motion segmentation. The measure was then generalised to cases involving 3D translational motions. To guarantee correct and successful segmentation, a set of sufficient conditions in terms of the required W_{2D} (denoted by $\widetilde{W}$) was developed, for cases involving both 2D or 3D translational motions, via extensive experiments using synthetic images. In addition, the proposed conditions were not significantly affected by the variation of camera parameters, direction of translations in Z direction, size and depth objects in motion. The relevance of these conditions to the problems encountered in real-image applications was demonstrated by using those conditions to correctly predict the outcome of different translational-motion segmentation scenarios. In practice, these conditions can be used as a performance predictor for translational-motion segmentation since the term W_{2D} can be estimated using obtainable scene parameters, i.e. the expected *inlier* ratio, the difference between direction of translations, the desired sensitivity of the system in terms of the amount of translation and the scale of noise. These conditions serve as a guideline for practitioners in designing motion-segmentation solutions for computer-vision applications.

Chapter 6

Analysis of Planar-Motion Segmentation

Many computer-vision applications involve 3D objects having planar motions. The motions in these applications are commonly assumed to be restricted to a single or multiple planes perpendicular to the camera optical axis. This is a common scenario because the distances between the camera and the objects, in those applications, are often much larger than the object motions along the camera optical axis [39, 70, 83, 121]. This chapter studies the feasibility of motion segmentation in a scene containing 3D-rigid objects having multiple planar motions. Assuming that the scene is viewed by an uncalibrated camera and the motion parameters are not known in advance, the most suitable motion model is the affine fundamental matrix [104].

The analysis starts by deriving the theoretical conditions to guarantee successful planar-motion segmentation using affine fundamental matrix in section 6.1. The validity of these conditions are examined via experiments using synthetic images (in

section 6.2) and their usefulness is demonstrated in experiments using real image data (in section 6.3). Finally section 6.4 concludes the chapter.

6.1 Segmentation of motion with affine fundamental matrix

To analyse the motion-segmentation problem, we consider a dynamic scene including two rigid 3D-objects with distinct planar motions and viewed by an uncalibrated camera. These two motions are denoted as motion-a and motion-b. Each motion consists of a rotation θ around the camera optical axis followed by a non-zero translation T i.e. θ_a and T_a for motion-a and θ_b and T_b for motion-b where $T_a = [T_{xa}\ T_{ya}\ 0]^\top$ and $T_b = [T_{xb}\ T_{yb}\ 0]^\top$. The aim of the analysis is to determine the theoretical limit of planar-motion segmentation — segmentation of points associated with motion-a from mixture of points having either motion-a and motion-b. Stationary points are not considered in this analysis, as the sufficient conditions for motion-background segmentation have already been established in chapter 4.

Similar to the previous analysis, we consider a point in 3D-space with coordinates $[X_i\ Y_i\ Z_i]^\top$ and denote its corresponding point in the image plane $\underline{m}_{1i} = [\underline{x}_{1i}\ \underline{y}_{1i}]^\top$, which moves to $\underline{m}_{2i} = [\underline{x}_{2i}\ \underline{y}_{2i}]^\top$ after a motion characterised by $(\theta\ T_x\ T_y)$. Then, using a camera with camera matrix A (in equation (2.2)), we have

$$\begin{aligned} \underline{x}_{1i} &= \frac{fX_i}{Z_i} + P_x, \quad \underline{x}_{2i} = \underline{x}_{1i}\cos\theta - \underline{y}_{1i}\sin\theta + \frac{fT_x}{Z_i} + \widetilde{P_x}, \\ \underline{y}_{1i} &= \frac{fY_i}{Z_i} + P_y, \quad \underline{y}_{2i} = \underline{x}_{1i}\sin\theta + \underline{y}_{1i}\cos\theta + \frac{fT_y}{Z_i} + \widetilde{P_y}, \end{aligned} \tag{6.1}$$

where:

$$\begin{aligned}\widetilde{P_x} =& P_x(1-\cos\theta) + P_y \sin\theta, \\ \widetilde{P_y} =& P_y(1-\cos\theta) - P_x \sin\theta.\end{aligned} \tag{6.2}$$

The symbols θ, T_x and T_y in equations (6.1) and (6.2) represent the motion parameters and the subscript a and b are used to identify the associated motion (motion-a and motion-b). The measured coordinates of points in the image plane are assumed to be contaminated by independently and identically-distributed (i.i.d) measurement noise e having a Gaussian distribution with zero mean and standard deviation σ_n:

$$\begin{aligned} x_{1i} &= \underline{x}_{1i} + e^1_{ix}, & y_{1i} &= \underline{y}_{1i} + e^1_{iy}, \\ x_{2i} &= \underline{x}_{2i} + e^2_{ix}, \quad \text{and} & y_{2i} &= \underline{y}_{2i} + e^2_{iy}. \end{aligned} \tag{6.3}$$

Without loss of generality, motion-a is considered as the target motion while motion-b is the unwanted one and, in terms of robust estimation, the matching points associated with motion-a are considered *inliers*, aimed to be separated from the points having motion-b, which are considered *outliers*.

The fundamental matrix of motion-a parameterised by θ_a, T_{xa} and T_{ya} is computed using equation (2.5) [3, 39, 124]

$$F_a = \frac{1}{f}\begin{bmatrix} 0 & 0 & T_{ya} \\ 0 & 0 & -T_{xa} \\ T_{xa}\sin\theta_a - T_{ya}\cos\theta_a & T_{ya}\sin\theta_a + T_{xa}\cos\theta_a & Q \end{bmatrix}, \tag{6.4}$$

where:

$$Q = (T_{ya}\cos\theta_a - T_{xa}\sin\theta_a - T_{ya})P_x + (T_{xa} - T_{ya}\sin\theta_a - T_{xa}\cos\theta_a)P_y. \tag{6.5}$$

Assuming that a robust estimator provides the true fundamental matrix given in (6.4), the Sampson distances d_i for all image points can be computed by substitution

of equations (6.1) and (6.4) in (3.4) [101, 104, 119], which yields:

$$d_i = \frac{1}{\sqrt{2(T_{ya}^2 + T_{xa}^2)}}[(\underline{x}_{1i} + e_{ix}^1)(T_{xa}\sin\theta_a - T_{ya}\cos\theta_a) + (\underline{y}_{1i} + e_{iy}^1)\cdots \\ (T_{ya}\sin\theta_a + T_{xa}\cos\theta_a) + (\underline{x}_{2i} + e_{ix}^2)T_{ya} - (\underline{y}_{2i} + e_{iy}^2)T_{xa} + Q]. \tag{6.6}$$

For the distances associated with points having motion-a, (denoted as d_{ai}) the expressions without noise terms in the above equation are equal to zero. This is because theoretically the numerator of the Sampson distance (in (3.4)) is zero for the target motion, according to equation (2.3). Thus, equation (6.6) can be simplified to

$$d_{ai} = \frac{1}{\sqrt{2(T_{ya}^2 + T_{xa}^2)}}[e_{ix}^1(T_{xa}\sin\theta_a - T_{ya}\cos\theta_a) + e_{iy}^1(T_{ya}\sin\theta_a + \cdots \\ T_{xa}\cos\theta_a) + e_{ix}^2 T_{ya} - e_{iy}^2 T_{xa}], \tag{6.7}$$

for points having motion-a. The distances given in (6.7) are a linear combinations of noise term e and the squared coefficients can be summed to one. Therefore, their distribution is identical to e, which is normally distributed with zero mean and standard deviation of σ_n i.e:

$$d_{ai} = e \approx N(0, \sigma_n^2). \tag{6.8}$$

Meanwhile, the Sampson distances associated with the points having motion-b (denoted by d_{bi}) can be expressed as

$$d_{bi} = \frac{1}{\sqrt{2(T_{ya}^2 + T_{xa}^2)}}[\underline{x}_{1i}(T_{xa}\sin\theta_a - T_{ya}\cos\theta_a) + \underline{y}_{1i}(T_{ya}\sin\theta_a + \cdots \\ T_{xa}\cos\theta_a) + \underline{x}_{2i}T_{ya} - \underline{y}_{2i}T_{xa} + Q] + e, \tag{6.9}$$

from equation (6.6). Combining the above equation with the world-to-image points relationship in (6.1) for points having motion-b, we obtain

$$d_{bi} = \frac{S\acute{\underline{x}}_{1i} + U\acute{\underline{y}}_{1i} + V}{\sqrt{2(T_{ya}^2 + T_{xa}^2)}} + e, \tag{6.10}$$

where the terms S, U and V are:

$$\begin{aligned} S &= T_{xa}(\sin\theta_a - \sin\theta_b) - T_{ya}(\cos\theta_a - \cos\theta_b), \\ U &= T_{xa}(\cos\theta_a - \cos\theta_b) + T_{ya}(\sin\theta_a - \sin\theta_b), \\ V &= \frac{f(T_{ya}T_{xb} - T_{xa}T_{yb})}{Z_{bi}}, \\ \acute{\underline{x}}_{1i} &= \underline{x}_{1i} - P_x, \qquad \text{and} \qquad \acute{\underline{y}}_{1i} = \underline{y}_{1i} - P_y. \end{aligned} \tag{6.11}$$

Note that, in the above equation, the subscript b is added to the term Z_i (Z_{bi}) to indicate that it is only associated with the depth of *outliers*, i.e. the points having motion-b. To simplify equation (6.10), all symbols associated with both motions are expressed in terms of its magnitude and direction — ϕ_a and ϕ_b denote the directions of T_a and T_b, respectively, where $T_{ya} = \|T_a\|\sin\phi_a$ and $T_{yb} = \|T_b\|\sin\phi_b$ — and rewritten, after algebraic manipulations using several trigonometric identities, as

$$d_{bi} = \sqrt{2}\sin\frac{\Delta\theta}{2}(\acute{\underline{x}}_{1i}\cos\Theta + \acute{\underline{y}}_{1i}\sin\Theta) + \frac{K_b}{Z_{bi}}\sin\Delta\phi + e, \tag{6.12}$$

where the symbols $\Delta\theta$, $\Delta\phi$, Θ and K_b are:

$$\begin{aligned} \Delta\theta &= \theta_a - \theta_b, \qquad \Delta\phi = \phi_a - \phi_b, \\ \Theta &= \phi_a - \frac{\theta_a + \theta_b}{2}, \qquad \text{and} \qquad K_b = \frac{f\|T_b\|}{\sqrt{2}}. \end{aligned} \tag{6.13}$$

The trigonometric terms in equation (6.12) are further simplified using the harmonic addition theorem [118] and expressed as

$$d_{bi} = G_{bi}\cos\breve{\Theta}_i\sin\frac{\Delta\theta}{2} + \frac{K_b}{Z_{bi}}\sin\Delta\phi + e, \tag{6.14}$$

where the term G_b and $\breve{\Theta}$ are:

$$\begin{aligned} G_{bi} &= \sqrt{2(\acute{\underline{x}}_{1i}^2 + \acute{\underline{y}}_{1i}^2)}, \quad \text{and} \\ \breve{\Theta}_i &= \Theta + \tan^{-1}(-\frac{\acute{\underline{x}}_{1i}}{\acute{\underline{y}}_{1i}}) + \left\{ \begin{array}{ll} 0 & \text{if } \acute{\underline{x}}_{1i} \geq 0 \\ \pi & \text{if } \acute{\underline{x}}_{1i} < 0 \end{array} \right\}. \end{aligned} \tag{6.15}$$

The term G_b in the above equations is associated with the location of points (having motion-b) appearing on the image plane and it is always positive. The values of $\cos\breve{\Theta}$ in equation (6.14) range from negative one to positive one based on the minimum and maximum values of the cosine term; thus, the range of the term $G_{bi}\cos\breve{\Theta}_i$ is

$$-\widehat{G_b} \leq G_{bi}\cos\breve{\Theta}_i \leq \widehat{G_b}, \tag{6.16}$$

where $\widehat{G_b}$ is the maximum value of G_{bi} depending on the locations of points associated with motion-b. Thus, the range of noise-free distances d_{bi} (denoted by $\underline{d}_{bi}$) associated with motion-b, from equations (6.14) and (6.16), can be expressed as:

$$-\widehat{G_b}|\sin\frac{\Delta\theta}{2}| + \frac{K_b}{Z_{bi}}\sin\Delta\phi \leq \underline{d}_{bi} \leq \widehat{G_b}|\sin\frac{\Delta\theta}{2}| + \frac{K_b}{Z_{bi}}\sin\Delta\phi. \tag{6.17}$$

Since the distribution of d_{ai} is the same as the distribution of noise term e, i.e. $N(0,\sigma_n^2)$ (as shown in equation (6.8)), the feasibility of identifying and segmenting points having motion-a depends on the distribution of distances associated with motion-b (d_{bi}). If both populations of distances d_{bi} and d_{ai} overlap each other, those motions would not be separable. Non-overlapping populations of distances (d_{bi} and d_{ai}) can ensure that those motions will be separable, as a robust estimator should be able to correctly segment all points having motion-a. Hence, in order to ensure that both populations of distances do not overlap, the following conditions must be satisfied

$$\underline{d}_{bi} \geq 5\sigma_n \quad \text{or} \quad \underline{d}_{bi} \leq -5\sigma_n, \tag{6.18}$$

where $\underline{d}_{bi}$ is the noise-free d_{bi} in equation (6.9). The condition in the above equation shows that the minimum or maximum value of d_{bi} needs to be far from the mean of

d_{ai} to minimise the possibility of overlapping distances and ensure successful segmentation. The threshold of $5\sigma_n$ in equation (6.18) is common in probability theory if the measurement noise is normally distributed (i.e. $N(0, \sigma_n^2)$) [22]. If this assumption is satisfied, theoretically only about 0.6% of d_{ai} and d_{bi} would overlap and at least 99.4% of points having motion-a will be correctly segmented [22]. In practice, the measurement values are always bounded and the above threshold would represent a perfect segmentation.

In order to link the condition for segmentation in (6.18) with the scene and motion parameters, the inequalities in (6.18) and (6.17) are combined and expressed as:

$$\widehat{G_b}|\sin\frac{\Delta\theta}{2}| \leq -5\sigma_n + \frac{K_b}{Z_{bi}}\sin\Delta\phi \quad \text{or} \quad \widehat{G_b}|\sin\frac{\Delta\theta}{2}| \leq -5\sigma_n - \frac{K_b}{Z_{bi}}\sin\Delta\phi. \tag{6.19}$$

The above inequalities can only be satisfied if the term $\frac{K_b}{Z_{bi}}\sin\Delta\phi$ or $-\frac{K_b}{Z_{bi}}\sin\Delta\phi$ is greater than or equal to $5\sigma_n$, since the term $\widehat{G_b}$ is always positive. In other words, one of the above conditions can only be satisfied if:

$$\frac{K_b}{Z_{bi}}|\sin\Delta\phi| \geq 5\sigma_n. \tag{6.20}$$

As such, the inequalities in equations (6.19) and (6.20) are expressed as:

$$\widehat{G_b}|\sin\frac{\Delta\theta}{2}| \leq \frac{K_b}{Z_{bi}}|\sin\Delta\phi| - 5\sigma_n. \tag{6.21}$$

Solving the inequalities in equations (6.21) and (6.20) for $\Delta\theta$ and $\Delta\phi$, the sufficient conditions for segmentation of motion-a are expressed as:

$$\begin{aligned} &|\frac{\Delta\theta}{2}| \leq \arcsin\frac{-5\sigma_n + \frac{K_b}{\bar{Z}_b}|\sin\Delta\phi|}{\widehat{G_b}} \qquad \text{and,} \\ &|\Delta\phi| \geq \arcsin\frac{5\sigma_n\bar{Z}_b}{K_b}. \end{aligned} \tag{6.22}$$

In the above equation, we assume that the term Z_{bi} (the depth of points associated with object having motion-b (*outliers*)) can be replaced by $\bar{Z}_b$, the average distance between the camera and the *outliers*. The justifications are twofold. First, in the target applications, we usually know the distance between the camera and the object in motion. For example, in a surveillance application, the distance between the camera and the surveillance area is known. Secondly, our experimental results, presented in the next section, show that the depth of the object having motion-b does not have a significant effect on the segmentation performance. It is important to note that, using the term $\bar{Z}_b$ to simplify Z_{bi} (in equation (6.22)) does not change the Sampson distances associated with the target motion (motion-a) and only affect the calculation of *outliers* residual (d_{bi} associated with motion-b). The distances (d_{ai}) associated with the points having motion-a are independent of their locations, as shown in (6.8).

In summary, we proposed the theoretical condition to guarantee successful segmentation (in (6.22)) of planar motion using affine fundamental matrix. The condition is based on obtainable motion and scene parameters i.e. the difference between rotation angles and translational directions ($\Delta\theta$ and $\Delta\phi$), the location of points having motion-b ($\widehat{G_b}$), the level of noise (σ_n) and the desired sensitivity of the system in terms of the amount of translation ($\frac{K_b}{Z_b}$). The validity of the above condition will be verified via experiments using synthetic images in section 6.2 and its usefulness as a performance predictor for planar-motion segmentation will be demonstrated using real-image data (in section 6.3).

6.2 Monte Carlo experiments using synthetic images

The Monte Carlo experiments for the analysis of planar-motion segmentation were divided into two parts. The first part of the experiments aimed to verify the theoretical conditions for successful segmentation of two planar motions in equation (6.22). The second part of the experiments was designed to examine the performance of the condition in terms of predicting the outcome of the segmentation under the variation of several scene parameters including *inlier* ratio ϵ (the ratio of the number of points having target motion, motion-a, over the total number of points, in equation (3.2)), size and depth of object associated with *outliers* (motion-b).

In each iteration in the Monte Carlo experiments, 2000 randomly generated points in the world-coordinate system having motion-a were mixed with the pairs of matching points undergoing motion-b. The number of points having motion-b was controlled by the value of ϵ in (3.2). The X and Y coordinates of the matching points having motion-b were randomly generated, while their Z coordinates were uniformly distributed according to $\bar{Z}_b \pm \frac{\delta_Z}{\bar{Z}_b}$ where $\frac{\delta_Z}{\bar{Z}_b} = 5\%$, 10% and 20%, representing the different depth (i.e. 10%, 20% and 40%) of object along the camera optical axis. All matching points having motion-a and b were projected on top of two images using a synthetic camera according to A_4

$$A_4 = \begin{bmatrix} 703 & 0 & 320 \\ 0 & 703 & 240 \\ 0 & 0 & 1 \end{bmatrix}, \tag{6.23}$$

representing a camera with field of view around 50°, focal length of 703 pixels, prin-

cipal point coordinate of [320 240] and image size of 640×480 pixels.

The motion and scene parameters in the experiments were based on two typical scenarios that are commonly encountered in real applications where the magnitude of $\|T_b\|$ and noise level σ_n are relatively small ($\frac{\|T_b\|}{\bar{Z}_b}$ is around 10% and σ_n is 0.5 or 1 pixel), and the locations of the points having motion-b are allowed to be in a wide region of the image ($\widehat{G_b} = 0.75G_{\max}$). The selected parameters for both scenarios are:

- Scenario-I with the parameters of $\frac{K_b}{\bar{Z}_b}$=50, $\sigma_n = 0.5$ and $\widehat{G_b} = 0.75G_{\max}$
- Scenario-II with the parameters of $\frac{K_b}{\bar{Z}_b}$=40, $\sigma_n = 1$ and $\widehat{G_b} = 0.75G_{\max}$

The values of $\frac{K_b}{\bar{Z}_b}$=50 and 40 in both scenarios correspond to $\|T_b\| = 1$m and 0.8m respectively, when average distance ($\bar{Z}_b$) between camera and object having motion-b is around 10 meter and camera matrix is according to A_4 in (6.23). Meanwhile, the term G_{bi} in (6.15) is at its maximum ($G_{\max}$) when $[\underline{\acute{x}}_{1i}\ \underline{\acute{y}}_{1i}]$=[320 240], i.e. the point located at each corner of the image, since the image size is 640×480 with the principal point at image center. Generally smaller values of $\widehat{G_b}$ correspond to the points associated with motion-b located closer to the principal point of the image.

The locations of points having motion-a could be anywhere in the image plane, since their distances are independent of their locations, as indicated by equations (6.7) and (6.8). However, the locations of points having motion-b are mandated by the value of $\widehat{G_b}$ and are assumed to be within $l\% \times l\%$ of the image, representing the size of object appearing on the image plane (in X and Y directions). This assumption is based on the fact that, in variety of computer-vision applications, the image/feature points associated with every moving object are largely confined to one part of an image. Concisely, by varying the values of l and $\frac{\delta_Z}{\bar{Z}_b}$, the size and the depth of object

having motion-b can be varied, along X, Y and Z axes.

The ground-truth matching points were perturbed by random noise assumed to be normally distributed $N(0, \sigma_n^2)$. The residuals for segmentation, in terms of Sampson distances d_i^2, associated with all points were calculated using equation (3.4) and the true fundamental matrix of target motion F_a. We assumed that an accurate estimate of F_a was provided by a robust estimator, hence it was calculated using equation (2.5) and given by equation (6.4). The segmentation was performed using the segmentation step of the MSSE [6] and again, the segmentation performance was measured by the ratio of the number of segmented points associated with motion-a over the true number of points having motion-a (denoted by ζ). Each experiment was repeated 1000 times and the statistical mean $\bar{\zeta}$ and standard deviation σ_ζ of 1000 ζs were recorded. Throughout the experiments several scene parameters were varied to examine their effect on the segmentation performance. The scene parameters included the *inlier* ratio (ϵ), the size (l) and depth ($\frac{\delta_Z}{Z_b}$) of object associated with *outliers* (motion-b). The pseudocode of the Monte Carlo experiments is shown in figure 6.1.

The first part of the experiment started with plotting of the theoretical conditions for segmentation of motion-a for Scenario-I and II using equation (6.22), and there are shown in figure 6.2. From our earlier analysis, the segmentation was expected to be successful when the values of $\Delta\theta$ and $\Delta\phi$ are from the white regions of figure 6.2, because in this region, the populations of d_i associated with both motions do not overlap. However, if $\Delta\theta$ and $\Delta\phi$ were outside the white regions in figure 6.2, there was no guarantee that points having motion-a would be successfully segmented. It is observed that the white regions in figure 6.2 are symmetrical, thus in the experiments we considered the range of $\Delta\theta$ and $\Delta\phi$ are between 0° to 90° with the conditions for

segmentation shown in figure 6.3.

Repeat ($\epsilon = 30\%$ to 80%) for Scenario-I and II.
Repeat ($\Delta\theta = 0^\circ$ to 90°) and ($\Delta\phi = 0^\circ$ to 90°).
Repeat ($l = 10\%$ to 30%) and ($\frac{\delta_Z}{\bar{Z}_b} = 5\%$ to 20%).

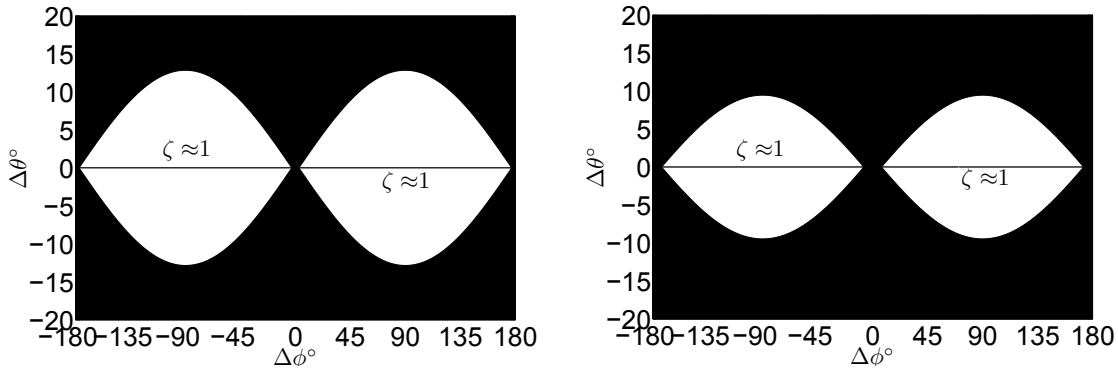
i. Repeat ($j = 1$ to 1000).
   1. Generate random θ_a, θ_b, T_a and T_b (between $\pm$ 180° and $\pm$ 1.5m) according to $\Delta\theta$, $\Delta\phi$, Scenario-I and II.
   2. Generate $N_i = 2000$ random pairs of points having motion-a according to θ_a and T_a.
   3. Generate N_o (3.2) random pairs of points (with uniformly distributed Z coordinates according to $\bar{Z}_b \pm \frac{\delta_Z}{\bar{Z}_b}$) having motion-$b$ according to θ_b, T_b and ϵ.
   4. Project all points on two images using a camera matrix A_4 in (6.23).
   5. Crop the points having motion-b so that they are within $l\% \times l\%$ of the image.
   6. Perturb all image points with Gaussian noise $N(0, \sigma_n^2)$.
   7. Calculate the true F_a of the points move having motion-a.
   8. Calculate the Sampson distances d_i of all points using the true F_a.
   9. Sort all d_i^2 and perform segmentation using MSSE.
   10. Record the ratio ζ (ratio of the segmented over the true number of points having motion-a).

ii. End.
iii. Calculate and record the mean and standard deviation of 1000 ζs.

End, End, End.

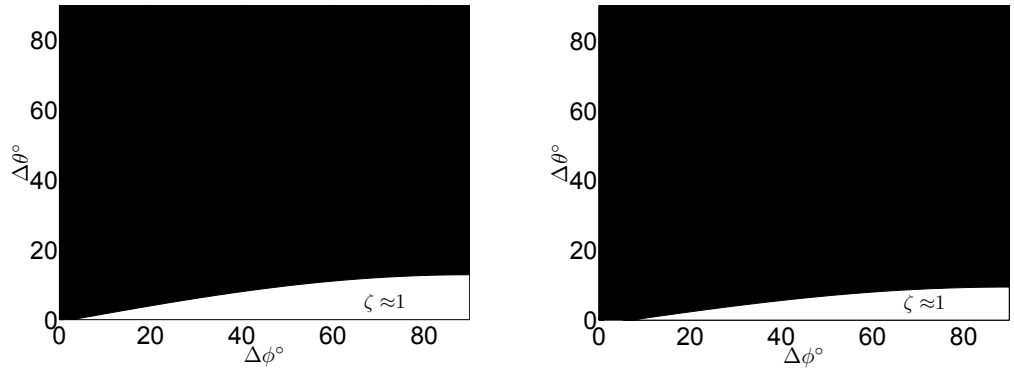
Figure 6.1: Pseudocode of Monte Carlo experiments for the analysis of planar-motion segmentation.

To verify the predictions in figure 6.3, we conducted thorough segmentation analysis for four different cases — where the values of $\Delta\theta$s and $\Delta\phi$s were selected from the white ($\Delta\theta = 6^\circ$ and $\Delta\phi = 50^\circ$ or $\Delta\theta = 10^\circ$ and $\Delta\phi = 80^\circ$) and black regions ($\Delta\theta = 2^\circ$ and $\Delta\phi = 5^\circ$ or $\Delta\theta = 4^\circ$ and $\Delta\phi = 10^\circ$) — of Scenario-I as shown in figure 6.3(a), by examining the segmentation performance ζ and the histogram of Sampson



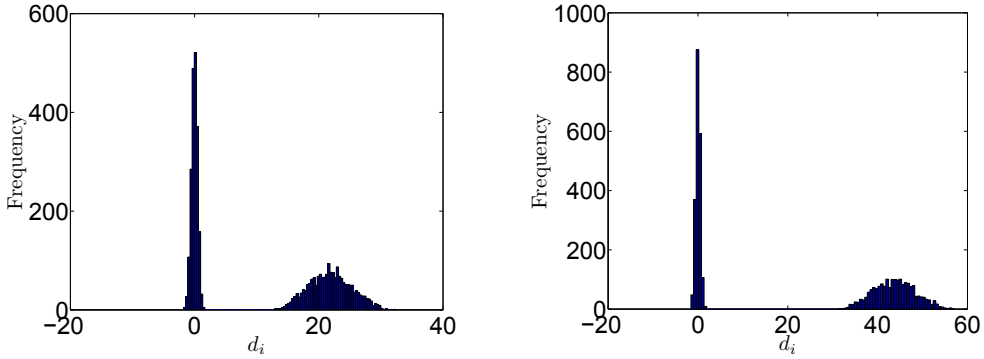
(a) Scenario-I (b) Scenario-II

Figure 6.2: Theoretical conditions for segmentation of Scenario-I and II.



(a) Scenario-I (b) Scenario-II

Figure 6.3: Theoretical conditions for segmentation of Scenario-I and II for $\Delta\theta$ and $\Delta\phi$ from 0° to 90°.

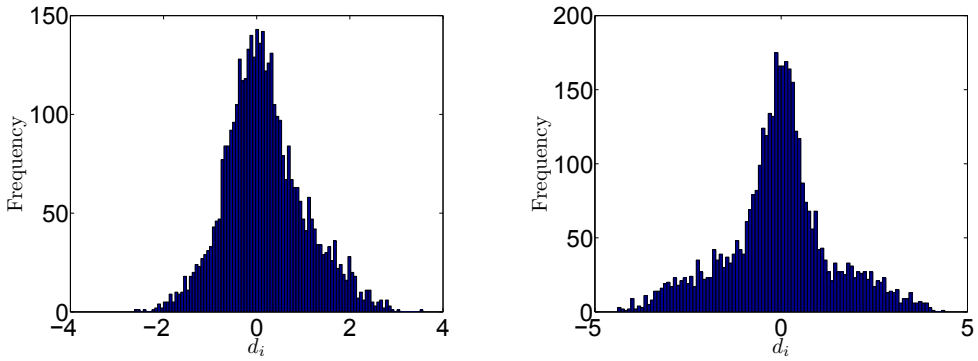


(a) $\Delta\theta = 6°$ and $\Delta\phi = 50°$, $\zeta = 0.98$ (b) $\Delta\theta = 10°$ and $\Delta\phi = 80°$, $\zeta = 0.98$

Figure 6.4: Histogram of d_i associated with all image points when $\Delta\theta$ and $\Delta\phi$ are from the white region of figure 6.3(a) (Scenario-I). The size of object having motion-b is according to $l = 20\%$ and depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$).

distances. The values of *inlier* ratio ϵ was around 50% while the locations and depth of object having motion-b were confined within $20\% \times 20\%$ ($l = 20\%$) of the image and 20% ($\frac{\delta_Z}{Z_b} = 10\%$), respectively.

The histogram of all distances and the associated segmentation performance ζs for all cases in Scenario-I are shown in figures 6.4 and 6.5. It can be observed in figure 6.4 that, when the values of $\Delta\theta$s and $\Delta\phi$s are selected from the white regions of figure 6.3(a), motion-a is correctly segmented as indicated by the values of ζ around one for both cases. As predicted, motion-a was successfully segmented because the populations of the distances associated with motion-a and motion-b did not overlap as shown in figure 6.4(a) and 6.4(b).



(a) $\Delta\theta = 2°$ and $\Delta\phi = 5°$, $\zeta = 1.91$ (b) $\Delta\theta = 4°$ and $\Delta\phi = 10°$, $\zeta = 1.85$

Figure 6.5: Histogram of d_i associated with all image points when $\Delta\theta$ and $\Delta\phi$ are from the black region of figure 6.3(a) (Scenario-I), The size of object having motion-b is according to $l = 20\%$ and depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$).

When the values of $\Delta\theta$ and $\Delta\phi$ were selected from the black region of figure 6.3(a), the values of ζ were larger than one indicating incorrect segmentation. As shown in figures 6.5(a) and 6.5(b), the failure to segment motion-a in these cases was due to the overlapping populations of distances associated with both motions.

The above results show that, in line with our theoretical predictions in equation (6.22), successful segmentation is guaranteed when the values of $\Delta\theta$ and $\Delta\phi$ are in the white regions of figure 6.3.

In the second part of the experiments, the effects of changing the following parameters to the sufficient conditions for segmentation — prescribed by equation (6.22) — of motions in Scenario-I and II were examined:

- the *inlier* ratio ϵ (from 30% to 80%),
- the size (from $l \times l = 10\% \times 10\%$ to $30\% \times 30\%$ of the image) of object having

motion-b, and

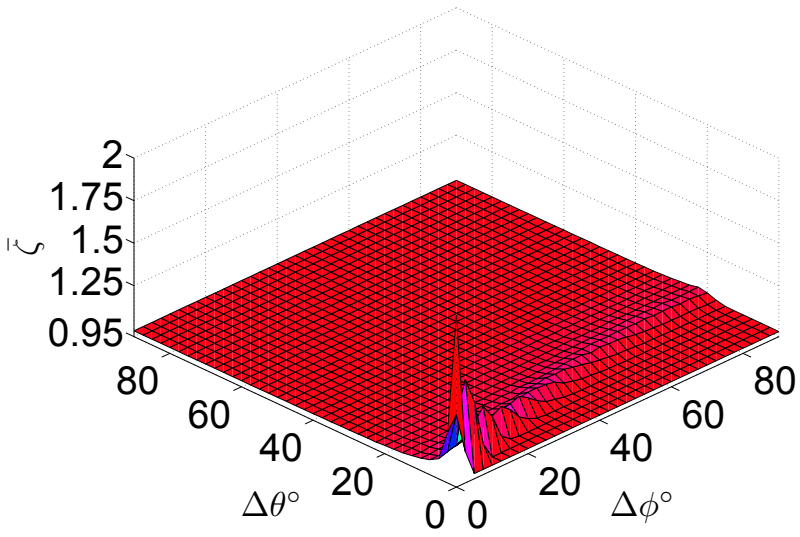
- the depth (from 10% to 40% i.e. $\frac{\delta_Z}{Z_b} = 5\%$ to 20%) of object having motion-b.

Note that the size, depth and location of the points associated with the object having motion-a, were randomly selected since their distances are independent of object-size, depth and location as shown in equations (6.7) and (6.8). The experiment was repeated 1000 times and the mean and standard deviation of ζs (denoted by $\bar{\zeta}$ and σ_ζ) were recorded for each pair of $\Delta\theta$ and $\Delta\phi$ and we again assumed that successful and consistent segmentation occurs when $\bar{\zeta} \approx 1$ and $\sigma_\zeta \leq 0.01$.

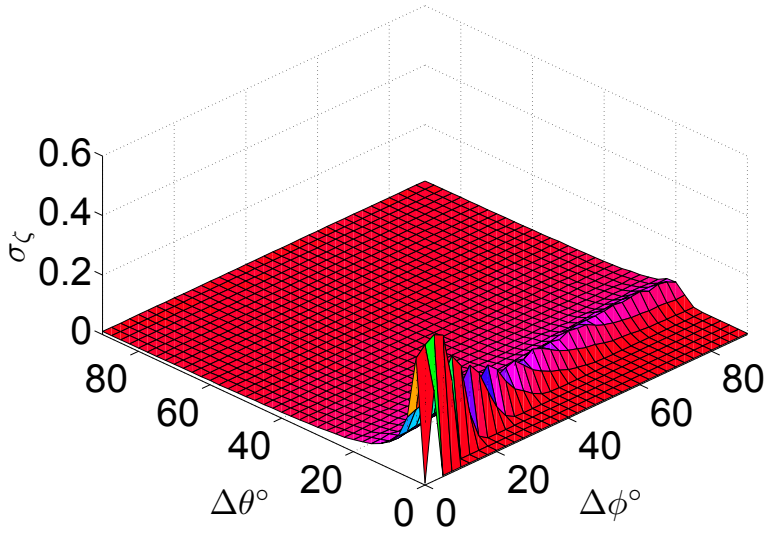
The values of $\bar{\zeta}$ and σ_ζ for all $\Delta\theta$s and $\Delta\phi$s are plotted in figures 6.6(a) and 6.6(b) when *inlier* ratio is 50% and the size of object having motion-b is 20% with depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$). These figures show that for small $\Delta\theta$ and $\Delta\phi$ (both $< 5°$), the segmentations were unsuccessful (denoted by $\bar{\zeta} > 1$). As both $\Delta\phi$ and $\Delta\theta$ were gradually increased toward 90°, both the values of $\bar{\zeta}$s and σ_ζs reduced to around one and zero respectively, which indicated successful and consistent segmentations.

Assuming that correct and consistent segmentation occurs when $\bar{\zeta} \approx 1$ and $\sigma_\zeta \leq 0.01$, the values of $\Delta\theta$ and $\Delta\phi$ for segmentation are extracted from figures 6.6(a) and 6.6(b) and represented by white regions in figure 6.7(b) when *inlier* ratio is 50% and the size of object having motion-b is 20% with depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$). A broad picture of the areas where the segmentations are correct and consistent for different values of *inlier* ratios, size and depth of object having motion-b is shown as white regions in figures 6.7 and 6.8.

Similar areas of white regions (areas where the segmentations were successful and consistent) are observed when the experimental results are compared with the

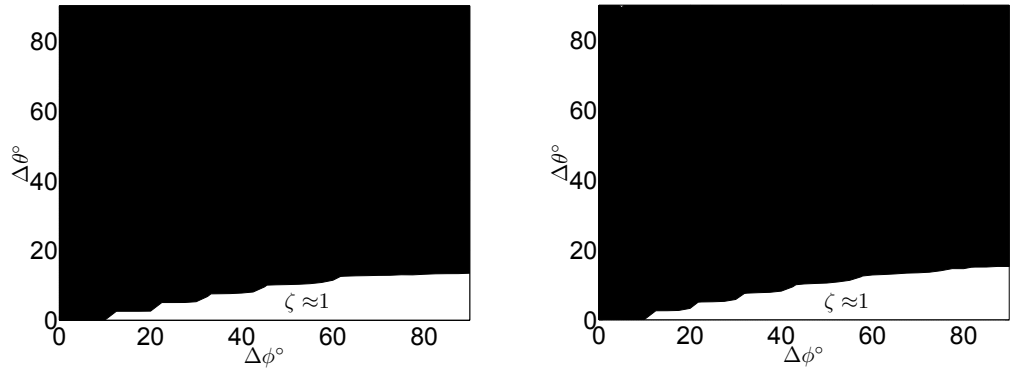


(a) $\bar{\zeta}$ vs $\Delta\theta$ and $\Delta\phi$



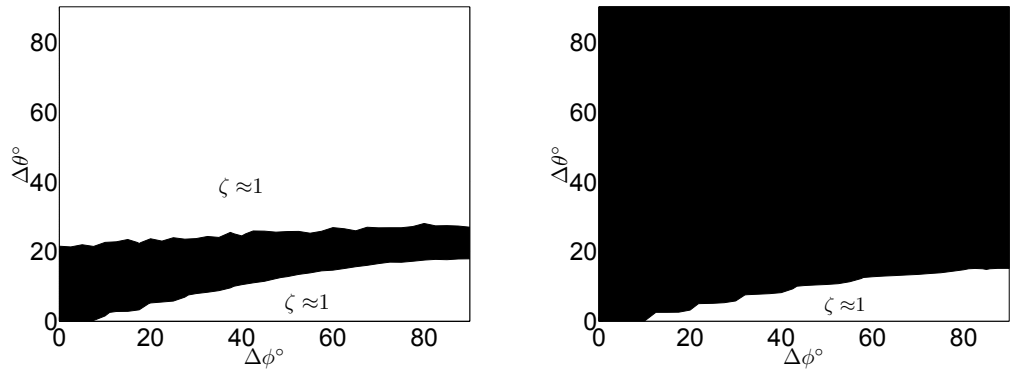
(b) σ_ζ vs $\Delta\theta$ and $\Delta\phi$

Figure 6.6: ζ vs $\Delta\theta$ and $\Delta\phi$ for Scenario-I when *inlier* ratio $\epsilon = 50\%$ and the size of object having motion-b is according to $l = 20\%$ and depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$).



(a) $\epsilon = 30\%$

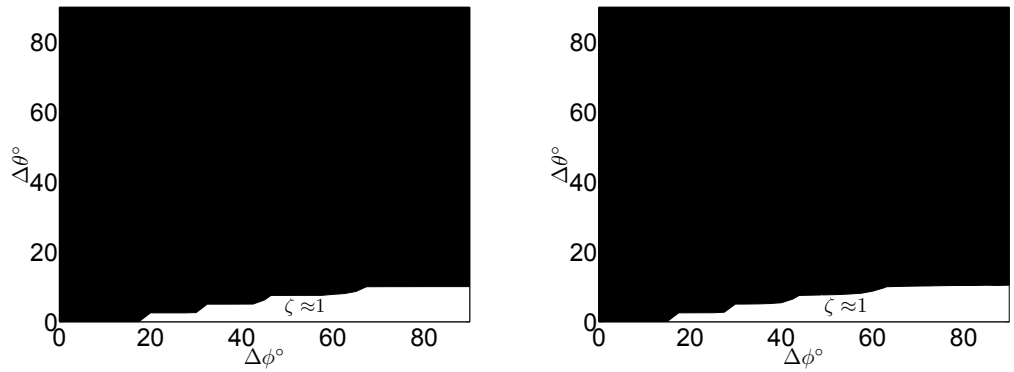
(b) $\epsilon = 50\%$



(c) $\epsilon = 80\%$

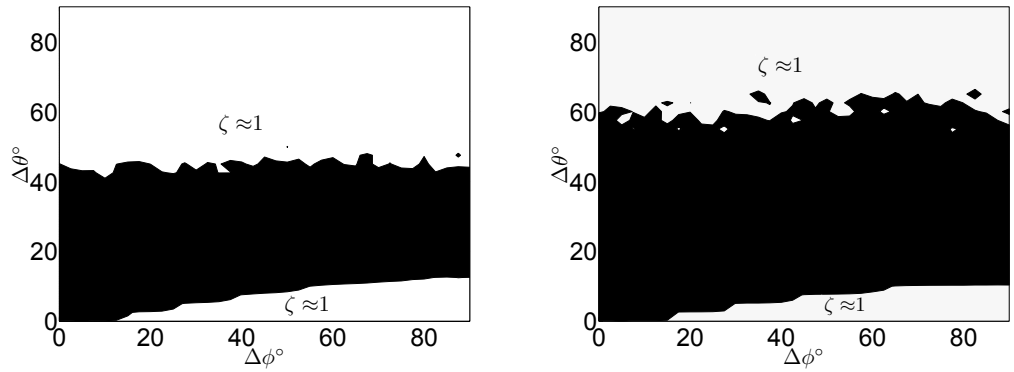
(d) $\epsilon = 30\%$ using larger object

Figure 6.7: Regions for successful segmentation (white region) for Scenario-I for various values of inlier ratio ϵ and size of object. Figures (a), (b) and (c) are the results when the size of object having motion-b is according to $l = 20\%$ and depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$). Figure (d) is the result when using larger object having motion-b with size ($l = 30\%$) and depth of 40% ($\frac{\delta_Z}{Z_b} = 20\%$).



(a) $\epsilon = 30\%$

(b) $\epsilon = 50\%$



(c) $\epsilon = 80\%$

(d) $\epsilon = 80\%$ using smaller object

Figure 6.8: Regions for successful segmentation (white region) for Scenario-II for various values of inlier ratio ϵ and size of object. Figures (a), (b) and (c) are the results when the size of object having motion-b is according to $l = 20\%$ and depth of 20% ($\frac{\delta_Z}{Z_b} = 10\%$). Figure (d) is the result when using smaller object having motion-b with size ($l = 10\%$) and depth of 10% ($\frac{\delta_Z}{Z_b} = 5\%$).

theoretical predictions for both scenarios — figures 6.7(a), 6.7(b) and 6.7(c) with 6.3(a) and figures 6.8(a), 6.8(b) and 6.8(c) with 6.3(b). In addition, when the *inlier* ratios are large (ϵ around 80%), motion-a was also successfully segmented when $\Delta\theta > 20^\circ$ (for Scenario-I in figure 6.7(c)) and $\Delta\theta > 45^\circ$ (for Scenario-II in figure 6.8(c)), indicating that the segmentations were relatively less challenging when relatively small number of points associated with motion-b were present.

To examine the effect of varying object size on the segmentation performance, the experiments were repeated for object (having motion-b) with larger size and depth ($l \times l = 30\% \times 30\%$ and 40% or $\frac{\delta_Z}{\bar{Z}_b} = 20\%$) and smaller size and depth ($l \times l = 10\% \times 10\%$ and 10% or $\frac{\delta_Z}{\bar{Z}_b} = 5\%$). The areas for successful segmentation of motion-a when the size and depth of object having motion-b are varied are shown as the white regions in figures 6.7(d) and 6.8(d). Comparison of the experimental results and the analytical predictions for both scenarios (i.e. figure 6.7(d) with 6.3(a) and 6.8(d) with figure 6.3(b)) indicates no significant variation to the predicted region for successful segmentation when the size and depth of object having motion-b (l and $\frac{\delta_Z}{\bar{Z}_b}$) were varied in a fairly broad range. These observations also support the validity of using the term $\bar{Z}_b$ (average distance between the camera and the *outliers* (object having motion-b)), in deriving the sufficient condition for segmentation (in 6.22), to represent the depth Z_{bi} of points associated with object having motion-b (in 6.19).

These results show the relevance of the theoretical conditions for segmentation in predicting the outcome of motion segmentation involving planar motions. If the inequalities in (6.22) are satisfied, the segmentation is guaranteed to be successful, irrespective of the *inlier* ratios, size and depth of objects in motion.

6.3 Experiments using real images

The usefulness of the derived conditions to predict the outcome of motion segmentations was further demonstrated via experiments using real-image data. In these experiments, we again considered a scene containing two objects having two different planar motions, motion-a and motion-b. The points associated with motion-a were considered as *inliers* aimed to be segmented from points having motion-b (*outliers*).

The experimental aim here is to investigate the theoretical limit of motion segmentation and how imperfect estimate of the fundamental matrix would affects the conditions for segmentation is beyond the scope of this work. In practice, the fundamental matrix can be accurately estimated using a number of robust methods [3, 101, 124] and the *gross outliers* can be removed by applying a robust estimator as part of the motion segmentation process [123]. The issues of estimation including estimating the fundamental matrix in terms of both the feasibility and the accuracy, have already been thoroughly analysed [42, 44]. As such in our experiments using real-image data — identical to our earlier theoretical analysis and Monte Carlo experiments — we assumed that an accurate estimate of the fundamental matrix of motion-a was provided by a robust estimator and there were no mismatches (*gross outliers*) in the image data. Thus, we calculated the fundamental matrix of motion-a using equation (2.5) and manually removed the *gross outliers* in the data. These assumptions needed to be taken in order to eliminate potential errors, from the estimation of the fundamental matrix and from the presence of *gross outliers* in the data, in the experimental results.

The experiments started with calibration of the camera using a camera-calibration

toolbox [16]. Two specially designed and fabricated triangular-shaped 3D objects — consisted of non-repeating patterns on each side of the object that was visible to the camera — were used to represent the objects having planar motions. The non-repeating patterns on both 3D objects ensured maximum number of image point can be extracted from their images and minimised the possibility of having *gross outliers* in the data. Each 3D object was moved according to either motion-a or motion-b and their images before and after each pair of motions (motion-a and motion-b) were acquired. The experiment was designed to represent a situation in Scenario-I; thus the scene parameters and location of the object having motion-b were selected such that the values of $\frac{K_b}{Z_b} = 50$ and $\widehat{G_b} = 0.75 G_{\max}$. In addition, the values of $\Delta\theta$ and $\Delta\phi$ for each pair of motion-a and motion-b were varied from 0° to 90°.

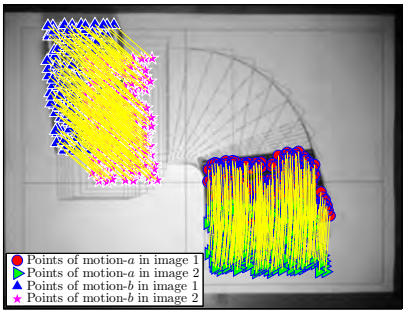
All corresponding feature points from all images were determined using the SIFT algorithm [58, 56] after the reduction of image distortion using radial and tangential-distortion models suggested by the camera-calibration toolbox [16]. Incorrect matches and static points from the background were manually eliminated. The *inlier* ratio ϵ in each pair of images was varied from 30% to 80% by removing some of the points having motion-b while maintaining all points having motion-a. The standard deviation of measurement noise σ_n throughout the experiment was around 0.5 pixel estimated from equation (3.5). Motion segmentation was performed using the squares of the Sampson-distance measure as the residuals, computed using the calculated fundamental matrix of motion-a. Again, we followed the segmentation steps of the MSSE [6] and the segmentation performance was measured by the ratio of segmented points over the true number of points having motion-a (denoted by the symbol ζ).

The experiments were conducted by performing segmentation for motions with

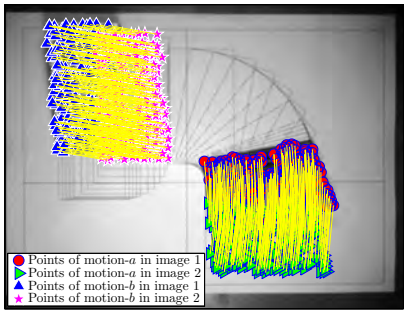
four different values of $\Delta\theta$ and $\Delta\phi$ selected from both white ($\Delta\theta = 6^\circ$ and $\Delta\phi = 50^\circ$ or $\Delta\theta = 10^\circ$ and $\Delta\phi = 80^\circ$) and black regions ($\Delta\theta = 2^\circ$ and $\Delta\phi = 5^\circ$ or $\Delta\theta = 4^\circ$ and $\Delta\phi = 10^\circ$) of the theoretical conditions for segmentation — prescribed by equation (6.22) — for Scenario-I, shown in figure 6.3(a). The outcome of the segmentation, when the values of $\Delta\theta$ and $\Delta\phi$ fall in the white region of figure 6.3(a), were predicted to be successful. Figures 6.9(c), 6.9(d) and the corresponding values of segmentation performance ζ of around one indicate that points having motion-a are successfully segmented. As predicted, the segmentations are successful because the populations of the distances d_i associated with both motions do not overlap and are easily distinguished from each other, as shown in figures 6.9(e) and 6.9(f), respectively.

When the values of $\Delta\theta$ and $\Delta\phi$ were selected from the black region of figure 6.3(a) ($\Delta\theta = 2^\circ$ and $\Delta\phi = 5^\circ$ or $\Delta\theta = 4^\circ$ and $\Delta\phi = 10^\circ$), the segmentation results are incorrect in both cases as shown in figures 6.10(c), 6.10(d) and indicated by the values of ζ larger than one. The failure to segment the target motion in both cases are due to the overlapping populations of distances associated with both motions, as evidenced by the histograms in figures 6.10(e) and 6.10(f).

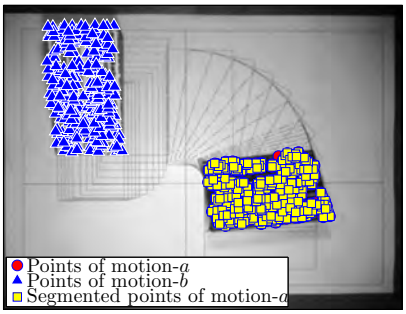
The segmentation performance (ζ) was plotted versus all values of $\Delta\theta$ and $\Delta\phi$ to examine its trend over variation of motion parameters. Figure 6.11 shows ζ versus $\Delta\theta$ and $\Delta\phi$ from our experiment using real-image data when *inlier* ratio is around 50%. It can be observed from figure 6.11 — similar to the earlier results from our Monte Carlo experiments in figure 6.6 — that the segmentation is not successful ($\zeta > 1$) when the angle differences $\Delta\theta$ and $\Delta\phi$ are small (i.e. both less than 5°) and the segmentation performance improves ($\zeta \approx 1$) when both $\Delta\theta$ and $\Delta\phi$ increase to around 90°.



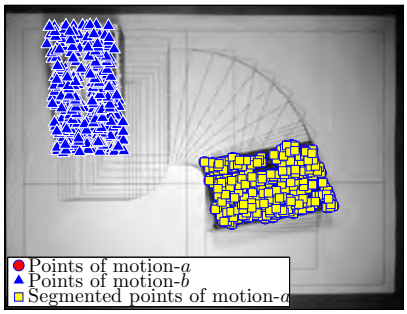
(a) White region, ($\Delta\theta = 6°$ and $\Delta\phi = 50°$)



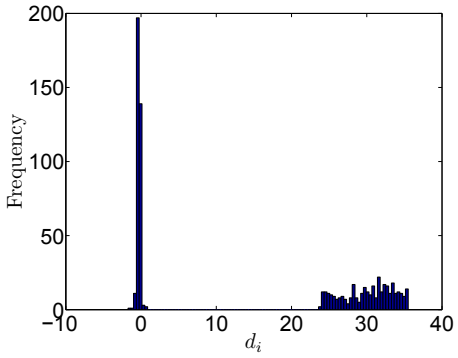
(b) White region, ($\Delta\theta = 10°$ and $\Delta\phi = 80°$)



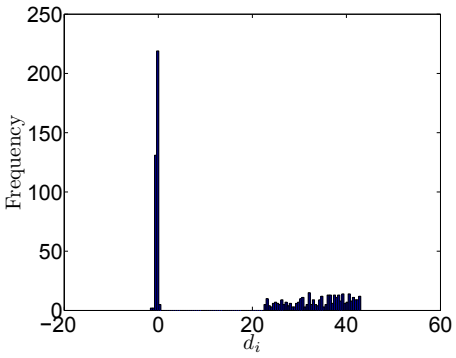
(c) $\Delta\theta = 6°$ and $\Delta\phi = 50°$, $\zeta = 0.99$



(d) $\Delta\theta = 6°$ and $\Delta\phi = 50°$, $\zeta = 0.99$

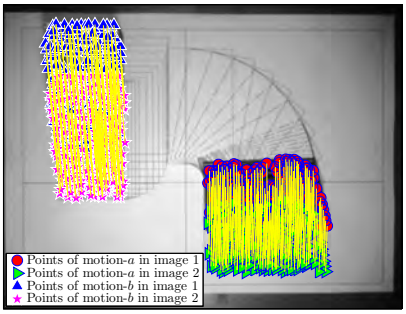


(e) White region, ($\Delta\theta = 6°$ and $\Delta\phi = 50°$)

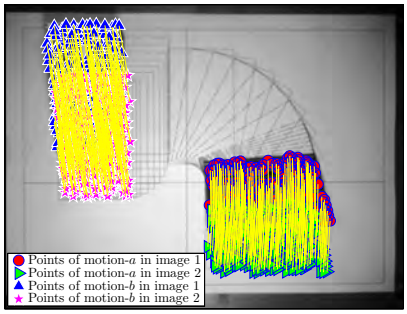


(f) White region, ($\Delta\theta = 6°$ and $\Delta\phi = 50°$)

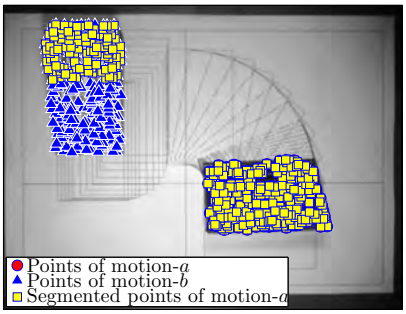
Figure 6.9: Segmentation results for motions with $\Delta\theta$ and $\Delta\phi$ in white region of figure 6.3(a). The ground-truth image points are superimposed onto the first image in (a) and (b) with $\epsilon = 50\%$, segmented points having motion-a in (c) and (d) and the histogram of d_i in (e) and (f).



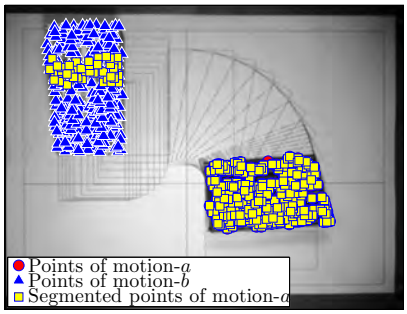
(a) Black region, ($\Delta\theta = 2°$ and $\Delta\phi = 5°$)



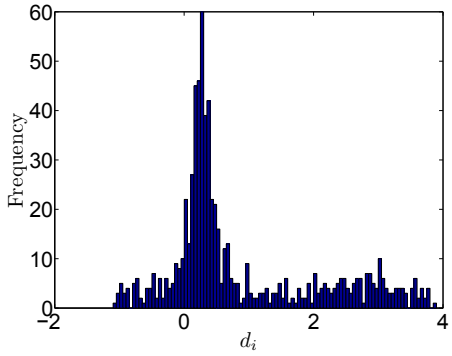
(b) Black region, ($\Delta\theta = 4°$ and $\Delta\phi = 10°$)



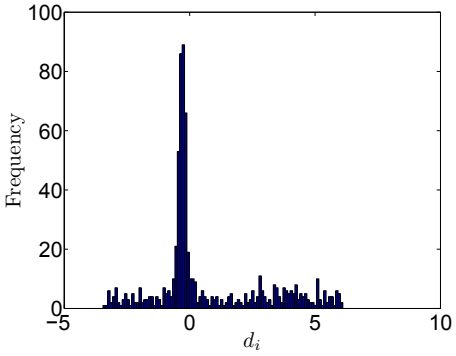
(c) $\Delta\theta = 2°$ and $\Delta\phi = 5°$, $\zeta = 1.42$



(d) $\Delta\theta = 4°$ and $\Delta\phi = 10°$, $\zeta = 1.16$

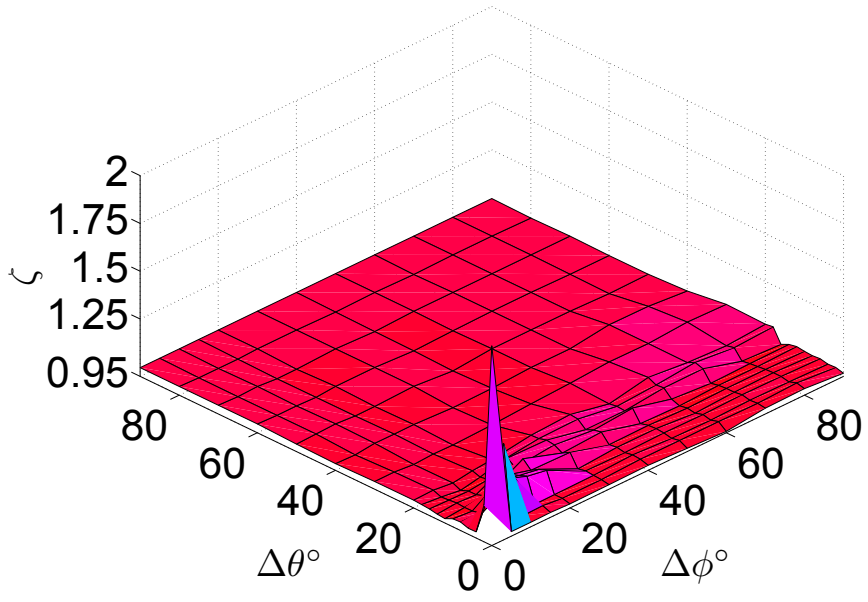


(e) Black region, ($\Delta\theta = 2°$ and $\Delta\phi = 5°$)



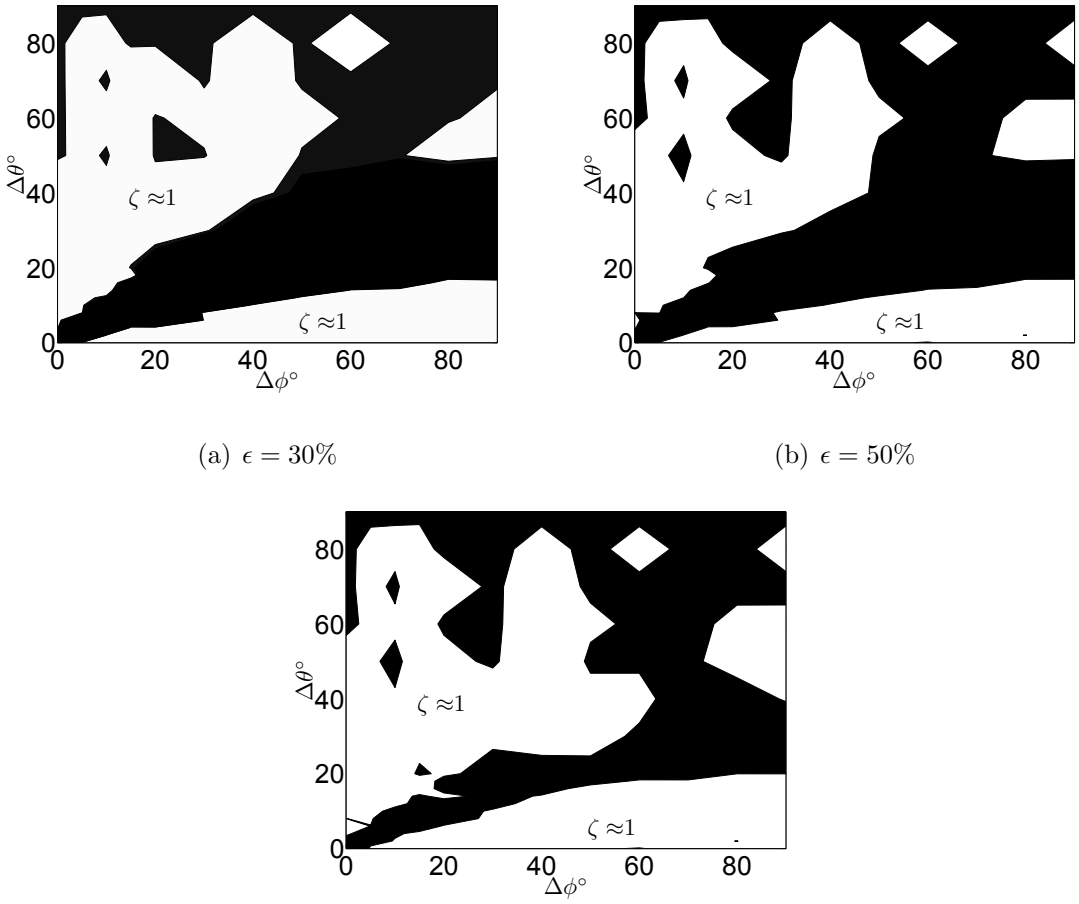
(f) Black region, ($\Delta\theta = 4°$ and $\Delta\phi = 10°$)

Figure 6.10: Segmentation results for motions with $\Delta\theta$ and $\Delta\phi$ in black region of figure 6.3(a). The ground-truth image points are superimposed onto the first image in (a) and (b) with $\epsilon = 50\%$, segmented points having motion-a in (c) and (d) and the histogram of d_i in (e) and (f).



(a) ζ vs $\Delta\theta$ and $\Delta\phi$

Figure 6.11: ζ vs $\Delta\theta$ and $\Delta\phi$ for Scenario-I when $\epsilon = 50\%$ from experiments using real-image data.



(a) $\epsilon = 30\%$

(b) $\epsilon = 50\%$

(c) $\epsilon = 80\%$

Figure 6.12: Regions for successful segmentation (white region) for Scenario-I for various ϵ from experiments using real-image data.

The white regions in figure 6.12 represent the corresponding values of $\Delta\theta$ and $\Delta\phi$ when the segmentation performance ζ are around one for all *inlier* ratios. We can observe the similarities of the white regions, where the segmentations are guaranteed to be successful, when the results from experiments using real and synthetic images were compared to the analytical condition for segmentation — figures 6.12(a)-6.7(a), figures 6.12(b)-6.7(b) and figures 6.12(c)-6.7(c), all with figure 6.3(a). In addition, we notice that the white regions in figure 6.12 appear larger when the *inlier* ratio is around 80%, since the segmentation is relatively less challenging and more feasible when *inlier* ratio is large.

These results demonstrate the capability of the derived conditions to correctly predict the outcome of motion segmentation involving planar motions. They also show that the theoretical derivation and the results of the Monte Carlo experiments are very relevant to the motion-segmentation problem in computer-vision applications.

6.4 Conclusion

Sufficient conditions for the segmentation of planar motions were theoretically derived to guarantee correct and successful segmentation. The validity of the derived conditions was examined via experiments using both synthetic and real images. The experimental results showed that those conditions were very relevant to the problems encountered in real-world applications and they were not significantly affected by variation of *inlier* ratio, size and depth of objects in motion. In practice, the derived conditions are capable of predicting the outcome of planar-motion segmentation, since they are expressed in terms of obtainable scene parameters as shown in equa-

tion (6.22), i.e. angle difference between both translational and rotational motions, location of points associated with objects having unwanted motion, estimated noise level and the desired sensitivity of the system in terms of the magnitude of translation. These conditions serve as guidelines for practitioners involved in designing computer-vision applications.

Chapter 7

Conclusions

In this work, we have considered the feasibility analysis of motion segmentation using the fundamental matrix motion model. The focus was on the motions of rigid 3D objects viewed by an uncalibrated camera. The quantitative measures for the degree of separation for motion-background, translational-motion and planar-motion segmentation were theoretically derived; and the feasibility of segmentation was expressed as a set of sufficient conditions for segmentation determined via extensive Monte Carlo experiments using synthetic images. The usability of these conditions were then demonstrated by experiments using real-image data to correctly predict the outcome of segmentation for different types of motions.

In summary, the contribution of this thesis for applications of motion segmentation are as follows:

1. We have shown that a pure translation is not separable from static points associated with the background, and that the success of motion-background segmentation using the fundamental matrix depends on the rotation angle of

that particular motion. Additionally, a set of sufficient conditions for motion-background segmentation has been proposed in terms of minimum required rotation angle.

2. We have theoretically derived a measure for the degree of separation (called W_{2D}) between two 2D translations and have proposed this measure to guarantee the success of translational-motion segmentation for cases involving both 2D and 3D translations. A set of sufficient conditions for translational-motion segmentation in terms of the minimum required W_{2D} has been established via extensive experiments using synthetic images.

3. We have theoretically determined the degree of separation between two planar motions of 3D objects in terms of their motion and scene parameters. To guarantee successful planar-motion segmentation, a set of sufficient conditions in terms of rotational and translational angles ($\Delta\theta$ and $\Delta\phi$) for a particular scene has been proposed.

4. We have designed the experiments using synthetic images based on the Monte Carlo statistical method to examine the validity of the degree of separation and to develop the conditions for segmentation for motion-background, translational-motion and planar-motion segmentation. The experiments have been used to analyse motion-segmentation performance when several scene and motion parameters were varied — including translational and rotational parameters, size and location of object, *inlier* ratio, measurement noise and camera parameters.

5. We have carried out several experiments using real-image data to demonstrate

the capability of the proposed conditions to correctly predict the outcomes of different segmentation scenarios. The experimental results show that these conditions are very relevant to the problems encountered in real-world applications.

In practice, the value of the degree of separation between two motions can be estimated using obtainable scene and motion parameters thus, the outcome of motion segmentation can be predicted using the derived separability conditions. As such these conditions serve as a guideline for practitioners designing motion-segmentation solutions for computer-vision problems. In the work carried out in this thesis, sufficient conditions for motion segmentation for different types of motions using the fundamental matrix were developed and some of the results were presented in peer reviewed publications [7, 8, 9, 10].

7.1 Future work

This thesis has initiated a relatively new line of research in the feasibility analysis of motion segmentation and there are still a number of interesting problems to be addressed. First, the derived feasibility for segmentation — motion-background, translational-motion and planar-motion segmentation in chapter 4, 5 and 6, respectively — involving 3D objects are all in terms of sufficient conditions. It would be interesting to study the feasibility of the development of necessary conditions for motion segmentation.

Secondly, the fundamental matrix associated with the target motion can be accurately estimated using many available robust methods [3, 101, 124]. However, a robust method is not perfect and occasionally it may produce inaccurate estimate of the fun-

damental matrix. As such, it would be desirable to study the effect of inaccuracies in estimation of the fundamental matrix on the conditions for segmentation.

Finally, while we consider the cases of motion-background, translational-motion and planar-motion segmentations, the more general motion includes translation and rotation along and around all axes in 3D-space. Thus, it would be useful to expand the theoretical analysis of planar-motion segmentation to include motion parameters along the camera optical axis (T_z, θ_x and θ_y). Particular attention would be on establishing an effective way to theoretically derive the impact of those motions (T_z, θ_x and θ_y) on the location of points in the image plane.

Bibliography

[1] J. Aggarwal and N. Nandhakumar. On the computation of motion from sequences of images-a review. *Proceedings of the IEEE*, 76(8):917–935, Aug. 1988.

[2] J. Y. Aloimonos. Perspective approximations. *Image and Vision Computing*, 8:179–192, August 1990.

[3] X. Armangu and J. Salvi. Overall view regarding fundamental matrix estimation. *Image and Vision Computing*, 21:205–220, 2003.

[4] A. Bab-Hadiashar and R. Hoseinnezhad. Bridging parameter and data spaces for fast robust estimation in computer vision. *Proceedings of the Digital Image Computing: Techniques and Applications DICTA*, 0:1–8, 2008.

[5] A. Bab-Hadiashar and D. Suter. Robust optic flow computation. *International Journal of Computer Vision IJCV*, 29:59–77, August 1998.

[6] A. Bab-Hadiashar and D. Suter. Robust segmentation of visual data using ranked unbiased scale estimate. *Robotica*, 17(6):649–660, 1999.

[7] S. N. Basah, A. Bab-Hadiashar, and R. Hoseinnezhad. Conditions for motion-background segmentation using fundamental matrix. *IET Computer Vision*, 3(4):189–200, 2009.

[8] S. N. Basah, A. Bab-Hadiashar, and R. Hoseinnezhad. Conditions for segmentation of 2D translations of 3D objects. In *Proceedings of the International Conference on Image Analysis and Processing ICIAP*, volume 5716 LNCS, pages 82–91, Berlin, Heidelberg, 2009. Springer-Verlag.

[9] S. N. Basah, R. Hoseinnezhad, and A. Bab-Hadiashar. Limits of motion-background segmentation using fundamental matrix estimation. In *Proceedings of the Digital Image Computing: Techniques and Applications DICTA*, pages 250–256, Los Alamitos, CA, USA, 2008. IEEE Computer Society.

[10] S. N. Basah, R. Hoseinnezhad, and A. Bab-Hadiashar. Conditions for segmentation of motion with affine fundamental matrix. In *Proceedings of the International Symposium on Visual Computing ISVC*, volume 5875 LNCS, pages I: 415–424, Berlin, Heidelberg, 2009. Springer-Verlag.

[11] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding CVIU*, 110:346–359, June 2008.

[12] P. A. Beardsley, P. H. S. Torr, and A. Zisserman. 3D model acquisition from extended image sequences. In *Proceedings of the European Conference on Computer Vision ECCV*, volume II, pages 683–695, London, UK, 1996. Springer-Verlag.

[13] P. A. Beardsley and A. Zisserman. Affine calibration of mobile vehicles. In Mohr, R. and Chengke, W., editors, *Europe-China Workshop on Geometrical Modelling and Invariants for Computer Vision*, pages 214–221, Xi'an, China, 1995. Xidan University Press/Springer-Verlag.

[14] P. A. Beardsley, A. Zisserman, and D. W. Murray. Sequential updating of projective and affine structure from motion. *International Journal of Computer Vision IJCV*, 23(3):235–259, June 1997.

[15] S. S. Beauchemin and J. L. Barron. The computation of optical flow. *ACM Computing Surveys*, 27:433–466, September 1995.

[16] J. Bouguet. Camera calibration toolbox for matlab. *http://www.vision.caltech.edu/bouguetj/calib_doc/index.html*.

[17] S. Carlsson. Duality of reconstruction and positioning from projective views. In *Proceedings of the IEEE Workshop on Representation of Visual Scenes*, pages 85–82, Washington, DC, USA, 1995. IEEE Computer Society.

[18] H. Chen and P. Meer. Robust regression with projection based m-estimators. In *Proceedings of the International Conference on Computer Vision ICCV*, pages 878–885, 2003.

[19] H. Chen, P. Meer, and D. E. Tyler. Robust regression for data with multiple structures. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, I:1069–1075, 2001.

[20] L. Chen, Z. Wang, and Y. Jia. Stereo vision based floor plane extraction and camera pose estimation. In M. Xie, Y. Xiong, C. Xiong, H. Liu, and Z. Hu, editors, *Intelligent Robotics and Applications*, volume 5928 LNCS, pages 834–845. Springer Berlin / Heidelberg, 2009.

[21] O. Chum and J. Matas. Optimal randomized ransac. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 30:1472–1482, August 2008.

[22] M. Evans, N. Hastings, and B. Peacock. *Statistical Distributions*. Wiley, third edition, 2000.

[23] O. Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, Cambridge, MA, USA, 1993.

[24] O. Faugeras. *Three-dimensional computer vision: a geometric viewpoint*. MIT Press, Cambridge, MA, USA, 1993.

[25] O. Faugeras, Q.-T. Luong, and T. Papadopoulou. *The Geometry of Multiple Images: The Laws That Govern The Formation of Images of A Scene and Some of Their Applications*. MIT Press, Cambridge, MA, USA, 2001.

[26] O. D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig. In *Proceedings of the European Conference on Computer Vision ECCV*, pages 563–578. Springer-Verlag, 1992.

[27] O. D. Faugeras, Q.-T. Luong, and S. J. Maybank. Camera self-calibration: Theory and experiments. In *Proceedings of the European Conference on Computer Vision ECCV*, volume 588 LNCS, pages 321–334. Springer, 1992.

[28] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24:381–395, June 1981.

[29] G. H. Golub and C. F. Van Loan. *Matrix computations (3rd ed.)*. Johns Hopkins University Press, Baltimore, MD, USA, 1996.

[30] L. Goshen and I. Shimshoni. Balanced exploration and exploitation model search for efficient epipolar geometry estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 30:1230–1242, July 2008.

[31] R. Gupta and R. Hartley. Linear pushbroom cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 19:963–975, September 1997.

[32] C. Harris and M. Stephens. A Combined Corner and Edge Detection. In *Proceedings of The Fourth Alvey Vision Conference*, pages 147–151, 1988.

[33] R. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *Proceedings of the European Conference on Computer Vision ECCV*, pages 579–587, London, UK, 1992. Springer-Verlag.

[34] R. Hartley. Euclidean reconstruction from uncalibrated views. In *Applications of Invariance in Computer Vision*, pages 237–256. Springer-Verlag, 1993.

[35] R. Hartley. Projective reconstruction and invariants from multiple images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 16:1036–1041, October 1994.

[36] R. Hartley. In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 19:580–593, June 1997.

[37] R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, pages 761–764. IEEE Comput. Soc. Press, 1992.

[38] R. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding CVIU*, 68(2):146–157, November 1997.

[39] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, second edition, 2003.

[40] R. Hesami, A. Bab Hadiashar, and R. Hoseinnezhad. A novel hierarchical technique for range segmentation of large building exteriors. In *Proceedings of the International Symposium on Visual Computing ISVC*, volume 4842 LNCS, pages II: 75–85. Springer, 2007.

[41] P. W. Holland and R. E. Welsch. Robust regression using iteratively reweighted least-squares. *Communications in Statistics - Theory and Methods*, 6:813–827, 1977.

[42] R. Hoseinnezhad and A. Bab-Hadiashar. Consistency of robust estimators in multi-structural visual data segmentation. *Pattern Recognition*, 40(12):3677–3690, 2007.

[43] R. Hoseinnezhad and A. Bab-Hadiashar. A novel high breakdown m-estimator for visual data segmentation. *Proceedings of the International Conference on Computer Vision ICCV*, pages 1–6, Oct. 2007.

[44] R. Hoseinnezhad, A. Bab-Hadiashar, and D. Suter. Finite sample bias of robust estimators in segmentation of closely spaced structures: A comparative study. *Journal of Mathematical Imaging and Vision JMIV*, 37(1):66–84, 2010.

[45] R. Hoseinnezhad, B.-N. Vo, and D. Suter. Fast segmentation of multiple motions. In *Proceedings of the Cognitive systems with Interactive Sensors COGIS 2009*, Paris, France, 2009. SEE.

[46] P. Hough. Method and Means for Recognizing Complex Patterns. U.S. Patent 3.069.654, 1962.

[47] P. J. Huber. Robust regression: Asymptotics, conjectures and Monte Carlo. *Annals of Statistics*, 1:799–821, 1973.

[48] P. J. Huber. *Robust Statistics*. Wiley Series in Probability and Statistics. Wiley-Interscience, 1981.

[49] F. Jurie and C. Schmid. Scale-invariant shape features for recognition of object categories. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, volume II, pages 90–96, 2004.

[50] T. Kadir and M. Brady. Saliency, scale and image description. *International Journal of Computer Vision IJCV*, 45:83–105, November 2001.

[51] J. Klappstein, F. Stein, and U. Franke. Detectability of moving objects using correspondences over two and three frames. In *Proceedings of the DAGM Conference on Pattern Recognition*, pages 112–121, Berlin, Heidelberg, 2007. Springer-Verlag.

[52] P. Kovesi. Matlab and octave functions for computer vision and image analysis. *http://www.csse.uwa.edu.au/ pk/Research/MatlabFns/index.html*.

[53] K.-M. Lee, P. Meer, and R.-H. Park. Robust adaptive segmentation of range images. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 20:200–205, 1998.

[54] T. Lindeberg. Feature detection with automatic scale selection. *International Journal of Computer Vision IJCV*, 30:79–116, November 1998.

[55] H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293(5828):133–135, 1981.

[56] D. G. Lowe. Sift keypoint detector. *http://www.cs.ubc.ca/ lowe/keypoints/*.

[57] D. G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision ICCV*, volume 2, pages 1150–1157, Washington, DC, USA, 1999. IEEE Computer Society.

[58] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision IJCV*, 60:91–110, 2004.

[59] Q. T. Luong, R. Deriche, O. Faugeras, and T. Papadopoulo. On determining the fundamental matrix : analysis of different methods and experimental results. Technical Report RR-1894, 1993.

[60] Q.-T. Luong and O. Faugeras. The fundamental matrix: theory, algorithms, and stability analysis. *International Journal of Computer Vision IJCV*, 17:43–75, 1995.

[61] Y. Ma, S. Soatto, J. Kosecka, and S. S. Sastry. *An Invitation to 3-D Vision: From Images to Geometric Models*. SpringerVerlag, 2003.

[62] S. Maybank. *Theory of Reconstruction from Image Motion*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1992.

[63] R. Mech and M. Wollborn. A noise robust method for 2D shape estimation of moving objects in video sequences considering a moving camera. *Signal Processing, Vol. 66, No. 2, pp. 203-217*, 0(0), 1998.

[64] G. Medioni and S. B. Kang. *Emerging Topics in Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA, 2004.

[65] P. Meer, D. Mintz, A. Rosenfeld, and D. Y. Kim. Robust regression methods for computer vision: a review. *International Journal of Computer Vision IJCV*, 6:59–70, April 1991.

[66] K. Mikolajczyk and C. Schmid. Indexing based on scale invariant interest points. In *Proceedings of the International Conference on Computer Vision ICCV*, volume 1, pages 525–531, 2001.

[67] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision IJCV*, 60:63–86, October 2004.

[68] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 27:1615–1630, October 2005.

[69] J. V. Miller and C. V. Stewart. Muse: Robust surface fitting using unbiased scale estimates. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, pages 300–306, Washington, DC, USA, 1996. IEEE Computer Society.

[70] J. L. Mundy and A. Zisserman, editors. *Geometric invariance in computer vision*. MIT Press, Cambridge, MA, USA, 1992.

[71] D. Nistér. Preemptive ransac for live structure and motion estimation. *Machine Vision and Applications*, 16:321–329, 2005.

[72] M. Nixon and A. S. Aguado. *Feature Extraction & Image Processing, Second Edition*. Academic Press, January 2008.

[73] E. P. Ong and M. Spann. Robust optical flow computation based on least-median-of-squares regression. *International Journal of Computer Vision IJCV*, 31:51–82, February 1999.

[74] P. L. Rosin. Robust pose estimation. *IEEE Transactions on Systems, Man, and Cybernetics - Part B: Cybernetics*, 29(2):297–303, 1999.

[75] G. Roth and M. Levine. Segmentation of geometric signals using robust fitting. In *Proceedings of the International Conference on Pattern Recognition ICPR*, volume 2, pages 826–831, 1990.

[76] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

[77] P. J. Rousseeuw and A. M. Leroy. *Robust regression and outlier detection*. John Wiley & Sons, Inc., New York, NY, USA, 1987.

[78] K. Schindler. Source codes and datasets from Image Understanding Group, Department of Computer Science, TU Darmstadt. *http://www.iu.tu-darmstadt.de/datasets*.

[79] K. Schindler, U. James, and H. Wang. Perspective n-view multibody structure-and-motion through model selection. In *Proceedings of the European Conference on Computer Vision ECCV*, volume 3951 LNCS, pages 606–619, 2006.

[80] K. Schindler and D. Suter. Two-view multibody structure-and-motion with outliers through model selection. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 28(6):983–995, 2006.

[81] K. Schindler, D. Suter, and H. Wang. A model-selection framework for multibody structure-and-motion of image sequences. *International Journal of Computer Vision IJCV*, 79(2):159–177, 2008.

[82] L. S. Shapiro, A. Zisserman, and M. Brady. Motion from point matches using affine epipolar geometry. In J.-O. Eklundh, editor, *Proceedings of the European Conference on Computer Vision ECCV*, volume II, pages 73–84. Springer, 1994.

[83] L. S. Shapiro, A. Zisserman, and M. Brady. 3D motion recovery via affine epipolar geometry. *International Journal of Computer Vision IJCV*, 16(2):147–182, October 1995.

[84] H.-Y. Shum and R. Szeliski. Systems and Experiment Paper: Construction of Panoramic Image Mosaics with Global and Local Alignment. *International Journal of Computer Vision IJCV*, 36(2):101–130, February 2000.

[85] C. V. Stewart. Minpran: A new robust estimator for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 17:925–938, October 1995.

[86] C. V. Stewart. Bias in robust estimation caused by discontinuities and multiple structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 19:818–833, 1997.

[87] C. V. Stewart. Robust parameter estimation in computer vision. *Society for Industrial and Applied Mathematics Review (SIREV)*, 41:513–537, September 1999.

[88] C. Stiller and J. Konrad. Estimating motion in image sequences: A tutorial on modeling and computation of 2D motion. *IEEE Signal Processing Magazine*, 16:70–91, 1999.

[89] C. Stiller, J. Konrad, and R. Bosch. Estimating motion in image sequences - a tutorial on modeling and computation of 2D motion. *IEEE Signal Processing Magazine*, 16:70–91, 1999.

[90] R. Subbarao and P. Meer. Heteroscedastic projection based m-estimators. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR - Workshop*, volume III, pages 38–44, Washington, DC, USA, 2005. IEEE Computer Society.

[91] R. Subbarao and P. Meer. Beyond ransac: User independent robust regression. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR - Workshop*, pages 101–108, Washington, DC, USA, 2006. IEEE Computer Society.

[92] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography: a factorization method. *International Journal of Computer Vision IJCV*, 9:137–154, November 1992.

[93] B. Tordoff and D. W. Murray. Guided sampling and consensus for motion estimation. In *Proceedings of the European Conference on Computer Vision ECCV*, volume I, pages 82–98, London, UK, 2002. Springer-Verlag.

[94] P. H. S. Torr. *Motion Segmentation and Outlier Detection*. Phd thesis, Department of Engineering Science, University of Oxford, 1995.

[95] P. H. S. Torr. Geometric motion segmentation and model selection. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 356(1740):1321–1340, 1998.

[96] P. H. S. Torr, P. Beardsley, and D. W. Murray. Robust vision. In *British Machine Vision Conf.*, pages 145–154, UK, 1994.

[97] P. H. S. Torr and C. Davidson. Impsac: Synthesis of importance sampling and random sample consensus. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 25:354–364, March 2003.

[98] P. H. S. Torr, A. W. Fitzgibbon, and A. Zisserman. The problem of degeneracy in structure and motion recovery from uncalibrated image sequences. *International Journal of Computer Vision IJCV*, 32:27–44, August 1999.

[99] P. H. S. Torr and D. W. Murray. Statistical detection of independent movement from a moving camera. *Image and Vision Computing*, 11(4):180 – 187, 1993.

[100] P. H. S. Torr and D. W. Murray. Stochastic motion clustering. In *Proceedings of the European Conference on Computer Vision ECCV*, volume II, pages 328–337, Secaucus, NJ, USA, 1994. Springer-Verlag New York, Inc.

[101] P. H. S. Torr and D. W. Murray. The development and comparison of robust methods for estimating the fundamental matrix. *International Journal of Computer Vision IJCV*, 24(3):271–300, 1997.

[102] P. H. S. Torr, R. Szeliski, and P. Anandan. An integrated bayesian approach to layer extraction from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 23(3):297–303, 2001.

[103] P. H. S. Torr and A. Zisserman. Mlesac: a new robust estimator with application to estimating image geometry. *Computer Vision and Image Understanding CVIU*, 78:138–156, 2000.

[104] P. H. S. Torr, A. Zisserman, and S. J. Maybank. Robust detection of degenerate configurations while estimating the fundamental matrix. *Computer Vision and Image Understanding CVIU*, 71(3):312–333, 1998.

[105] P. H. S. Torr, A. Zisserman, and D. W. Murray. Motion clustering using the trilinear constraint over three views. In Mohr, R. and Chengke, W., editors, *Europe-China Workshop on Geometrical Modelling and Invariants for Computer Vision*, pages 118–125, Xi'an, China, 1995. Xidan University Press/Springer-Verlag.

[106] S. Valibeik and G.-Z. Yang. Segmentation and tracking for vision based human robot interaction. *Proceedings of the IEEE International Conference on Web Intelligence and Intelligent Agent Technology*, 3:471–476, 2008.

[107] R. Vidal and R. Hartley. Motion Segmentation with Missing Data using Power-Factorization and GPCA. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, volume II, pages 310–316, 2004.

[108] R. Vidal and Y. Ma. A unified algebraic approach to 2-d and 3-d motion segmentation and estimation. *Journal of Mathematical Imaging and Vision JMIV*, 25(3):403–421, 2006.

[109] R. Vidal, Y. Ma, and J. Piazzi. A new gpca algorithm for clustering subspaces by fitting, differentiating and dividing polynomials. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, pages I: 510–517, 2004.

[110] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (gpca). *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 27:1945–1959, 2005.

[111] R. Vidal, Y. Ma, S. Soatto, and S. Sastry. Two-view multibody structure from motion. *International Journal of Computer Vision IJCV*, 68:7–25, 2006.

[112] R. Vidal and S. Sastry. Optimal segmentation of dynamic scenes from two perspective views. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, volume II, pages 281–286, 2003.

[113] R. Vidal, S. Soatto, and S. Sastry. Segmentation of dynamic scenes from the multi-body fundamental matrix. In *Proceedings of the European Conference on Computer Vision ECCV - Workshop on Visual Modeling of Dynamic Scenes*, June 2002.

[114] T. Vieville and D. Lingrand. Using singular displacements for uncalibrated monocular visual systems. In *Proceedings of the European Conference on Computer Vision ECCV*, volume II, pages 207–216, London, UK, 1996. Springer-Verlag.

[115] H. Wang and D. Suter. Robust adaptive-scale parametric model estimation for computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 26(11):1459–1474, 2004.

[116] H. Wang and D. Suter. Robust fitting by adaptive-scale residual consensus. In *Proceedings of the European Conference on Computer Vision ECCV*, pages 107–118, 2004.

[117] D. Weinshall, M. Werman, and A. Shashua. Duality of multi-point and multi-frame geometry: Fundamental shape matrices and tensors. In *Proceedings of the European Conference on Computer Vision ECCV*, volume II, pages 217–227, London, UK, 1996. Springer-Verlag.

[118] E. W. Weisstein. Harmonic addition theorem. *From MathWorld-Wolfram Web Resource http://mathworld.wolfram.com/HarmonicAdditionTheorem.html*.

[119] J. Weng, T. Huang, and N. Ahuja. Motion and structure from two perspective views: Algorithms, error analysis, and error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 11(5):451–476, 1989.

[120] L. Wolf and A. Shashua. Two-body segmentation from two perspective views. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition CVPR*, volume I, pages 263–270, 2001.

[121] G. Xu and Z. Zhang. *Epipolar Geometry in Stereo, Motion, and Object Recognition: A Unified Approach*. Kluwer Academic Publishers, Norwell, MA, USA, 1996.

[122] X. Yu, T. D. Bui, and A. Krzyzak. Robust estimation for range image segmentation and reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI*, 16:530–538, 1994.

[123] D. Zhang and G. Lu. Segmentation of moving objects in image sequence: A review. *Circuits, Systems, and Signal Processing*, 20(2):143–183, 2001.

[124] Z. Zhang. Determining the epipolar geometry and its uncertainty: A review. *International Journal of Computer Vision IJCV*, 27(2):161–195, 1998.

[125] Z. Zhang and O. Faugeras. *3D Dynamic Scene Analysis*. Springer-Verlag, 1992.