The provision of this paper in an electronic form in this site is only for scholarly study purposes and any other use of this material is prohibited. What appears here is a near-publication draft of the final paper as appeared in the journal or conference proceedings. This is subject to the copyrights of the publishers. Please observe their copyrights.
Conditions for Motion-Background Segmentation Using
Fundamental Matrix

Shafriza Nisha Basah∗, Alireza Bab-Hadiashar†, and Reza Hoseinnezhad‡

1Faculty of Engineering and Industrial Science, Swinburne University of Technology, Victoria, Australia
2Department of Electrical & Electronic Engineering, The University of Melbourne, Victoria, Australia

Abstract

In common motion segmentation and estimation applications where the exact nature of objects’ motions and the camera parameters are not known a priori, the most general motion model (the fundamental matrix) is applied. Although the estimation of a fundamental matrix and its use for motion segmentation are well understood, the conditions governing the feasibility of segmentation for different types of motions are yet to be discovered. In this paper, we study the feasibility of separating motions of a 3D object from its static background using the fundamental matrix. We theoretically prove that a pure translational motion cannot be separated from its static background and the success of motion-background segmentation depends on the rotational part of the motion. An extensive set of controlled experiments using both synthetic and real images was conducted to validate the theoretical results. In addition, we quantified the conditions for successful motion-background segmentation in terms of the minimum required rotation angle. These results are useful for practitioners designing motion segmentation or estimation solutions for computer vision problems.

Keyword: Motion segmentation, Fundamental matrix, Robust estimator

∗sbasah@swin.edu.au
†abab-hadiashar@swin.edu.au
‡rezah@unimelb.edu.au
1 Introduction

An image sequence of a dynamic scene acquired by static or nearly static cameras contains corresponding features/points belong to either a moving object or the background. The points belonging to the background are static or have a very small motion. The background in many dynamic scenes may have textured areas therefore, a relatively large number of detected points may belong to the static (or nearly static) parts of the scene. As an example Fig.1 shows a sequence containing a moving object (3D object on the right) where the majority of the features detected by SIFT algorithm [1, 2] belong to the static background (book on the left).

![Figure 1](image1.png)  
(a)  

![Figure 1](image2.png)  
(b)

Figure 1: (a) and (b) Depiction of corresponding image points between two frames of a moving 3D object sequence generated by SIFT algorithm. Around 65% of the corresponding points belong to the static background.

Motion segmentation is the process of segmenting or recovering structure-and-motion (SaM) from images of a dynamic scene. The recovered SaM can be applied to various computer vision applications ranging from image rendering in multimedia applications to local navigation of autonomous robots. A motion segmentation algorithm usually consists of three main steps [3]:

1. Data primitives: determining the corresponding features/points belonging to the object in each image. Common types of data primitives are individual pixels, corners, lines, blocks, or regions [4]. In this study we use SIFT algorithm by Lowe [1, 2] to determine the corresponding image points in multiple images.

2. Motion model: deals with the representation of the motion which can be either 3D motion parameters or 2D optic flow.
3. Segmentation criteria: decide on the inlier-outlier dichotomy based on segmentation criteria or cost function.

There are many segmentation techniques proposed in the literature and they are classified based on their motion representation (2D optic flow or 3D motion parameters) or the way they perform the segmentation. A comprehensive classification of motion segmentation is provided in [3].

The main problem in motion segmentation and SaM recovery is that the exact nature of objects’ motions and the camera parameters are often not known a priori. Thus, the most general motion model which is the fundamental matrix is usually applied. The fundamental matrix also takes into account the 3D structure and real motion of the objects in space [3] and has the advantage of not requiring camera calibration. The acquisition of camera parameters from camera calibration is often impractical especially in the applications where the camera parameters are constantly changing even during the course of processing [5]. Motion segmentation and SaM recovery using the fundamental matrix are well understood and solved in the established work presented and summarised in (Chapters.9-12,[6]). Soon after that work, researchers resumed to the more challenging multibody structure-and-motion (termed MSaM by Schindler and Suter in [7]) where multiple objects in motions need to be concurrently estimated and segmented.

Well known examples of previous works in motion segmentation using the fundamental matrix are by Torr et al. [8, 9, 10, 11], Vidal et al. [12, 13] and Schindler and Suter [7]. Torr et al. presented a way to automatically determine the number of motions present in the scene and the type of motion model was appropriate for each motion to eliminate degeneracy (six motion models based on the fundamental matrix and homography were considered) [8, 11]. Then the parameters for each motion were determined by alternating between feature clustering and motion estimation in a probabilistic framework using Expectation Maximisation (EM) algorithm [8]. However the convergence of the EM algorithm highly depends on the initialisation procedures [10]. Vidal et al. proposed to estimate the number of moving objects in motion and cluster those motions using the multi-body fundamental matrix; the generalization of the epipolar constraint and the fundamental matrix of multiple motions [12]. This approach directly solves the motion parameters algebraically and eliminates the feature clustering stage which is an added advantage [13]. Schindler and Suter have implemented the geometric model selection to replace degenerate motion in dynamic scenes using the multi-body fundamental matrix [7]. However,
the conditions governing the feasibility of segmentation involving MSaM for different types of motions are yet to be established. These conditions are important as they provide information on the limits of current MSaM methods and would provide useful guidelines for practitioners designing motion segmentation or estimation solutions for computer vision problems.

The main focus of this paper is to study the feasibility of detection and segmentation of an unknown motion (from its static background) in an image sequence taken by an uncalibrated camera. It presents a theoretical analysis of separability of pure translational motions of 3D object from the static background. Then the required conditions for successful motion-background segmentation of 3D object are derived and investigated from both theoretical and experimental view points. In Section 2, we present a short review of motion segmentation and SaM recovery using the fundamental matrix. In Section 3, we show that a pure translational motion modeled by using the fundamental matrix is not separable from a static background (i.e. cannot be estimated or segmented). We then examine the required conditions for the feasibility of detection and segmentation of motions and provide quantitative measures for detectable motions in Section 4. Section 5 details the experiments using real images conducted to verify the proposed required conditions for segmentation in earlier sections and Section 6 concludes the paper. A preliminary shorter version of this paper has been presented at a conference [14]. Additional work in the current paper includes the method used to conduct the experiments using real images and extensive sets of experimental result using real images to verify the experimental results using synthetic images.

2 Motion segmentation using fundamental matrix

Consider a 3D object undergoing rotation and followed by a non-zero translation. The movement of the object is acquired by consecutive frames of a video sequence. The corresponding image points \( m_{1i} \) viewed in the first frame is transformed to \( m_{2i} \) in another frame where \( m_{1i} = [x_{1i}, y_{1i}, 1]^\top \) and \( m_{2i} = [x_{2i}, y_{2i}, 1]^\top \). These image points are related by:

\[
m_{2i}^\top F m_{1i} = 0
\]  

(1)
where $F$ is the $3 \times 3$ rank 2 fundamental matrix [6, 15, 16]. There are numerous techniques to compute (or estimate) $F$, including the linear, iterative and robust methods. For reviews of these techniques, readers are referred to [6, 15, 16, 17].

The video sequence generally contains measurement noise. In addition, mismatches are often generated during the step of determining corresponding image points. The measurement noise and mismatches cause uncertainty and errors to the segmentation result. Thus, robust estimators are usually applied to produce sufficiently accurate results. In general, most robust estimators, developed to solve this problem include three main steps [18]:

1. Optimization: returns an initial estimate of the motion parameters (fundamental matrix for the most general motion models) as a result of optimization of a cost function.

2. Segmentation: separates the inlier data points moving according to the target motion from others by processing their distances from the epipolar lines given by the fundamental matrix estimate.

3. Refinement: involves fine tuning of the fundamental matrix estimate by applying the least squares to the inliers detected in the segmentation step.

There are numerous robust estimators to perform the inlier-outlier separation in the segmentation step (recent examples of robust estimators with their own specific segmentation methods are Projection based M-Estimator (pbM) [19], Two-Step Scale Estimator (TSSE) [20] and Modified Selective Statistical Estimator (MSSE) [21]). The segmentation step of MSSE [21] is used in this paper because of its desired asymptotic and finite sample bias properties [22]. Although we use MSSE, the analysis is general and similar results will be obtained if other robust estimators are used.

The error measure is defined as a function of the distances such that it is minimum for the target object (inliers) and larger for the rest (outliers). Thus, segmentation is based on automatic separation of data points with small errors from the large ones. Four types of error measures for the estimation of fundamental matrix are:

1. The sum of squares of algebraic distances [11]:

$$D_r = \sum_i r_i^2 = \sum_i (m_{2i}^T F m_{1i})^2,$$

(2)
where the elements of $F$ are normalised.

2. The sum of squares of geometric distances [6]:

$$D_v = \sum_i d_{vi}^2,$$

$$d_{vi} = (x_{1i} - \hat{x}_{1i})^2 + (x_{2i} - \hat{x}_{2i})^2 + (y_{1i} - \hat{y}_{1i})^2 + (y_{2i} - \hat{y}_{2i})^2,$$

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 estimated fundamental matrix $\hat{F}$.

3. The sum of squares of Sampson distances [15, 11, 23]:

$$D = \sum_i d_i^2,$$

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

4. The sum of squares of Luong distances [24]:

$$D_L = \sum_i d_{Li}^2,$$

$$d_{Li} = \frac{m_{2i}^\top F m_{1i}}{\sqrt{\left[ \left( \frac{\partial}{\partial x_{1i}} \right)^2 + \left( \frac{\partial}{\partial y_{1i}} \right)^2 \right] m_{2i}^\top F m_{1i}}} + \frac{m_{2i}^\top F m_{1i}}{\sqrt{\left[ \left( \frac{\partial}{\partial x_{2i}} \right)^2 + \left( \frac{\partial}{\partial y_{2i}} \right)^2 \right] m_{2i}^\top F m_{1i}}}.$$

We adopted the Sampson distance measure because of two main reasons. Firstly, the Sampson distance provides the first order approximation (according to Torr and Murray [15] the accuracy of the approximation is around 4 or 5 significant figures) of the geometric distance and is computationally less expensive compared to the geometric distance [15, 11, 25]. Secondly, the Sampson distance measures often produce slightly better results compared to Luong distance [16].
3 Non-separability of pure translations

In general, a motion can be modeled as pure rotations around the origin followed by a pure translation. In this section, we study the separability of a pure translational motion from its background. The nature of this motion is unknown in advance, and therefore, it is modeled by the fundamental matrix. We aim to theoretically prove that a pure translational motion is not separable from its background if it is modeled by the fundamental matrix.

Consider a pure translation in X-Y plane (i.e. zero motion towards or away from the camera) denoted by $T = [T_x, T_y, T_z]^{\top}$ with $T_z = 0$. A dataset of $n = n_i + n_o$ pairs of matching points in two images is used for motion segmentation, with each point in images 1 and 2 denoted by $x_{1i}$, $y_{1i}$ and $x_{2i}$, $y_{2i}$. Points $[X_i, Y_i, Z_i]^{\top}$ with $i = 0, 1, 2 \ldots n_i$ denote the points belonging to the object undergoing translation $T$ while $i = n_i + 1, n_i + 2 \ldots n$ belong to static points representing the background of a dynamic scene. All points in the image plane are contaminated by measurement noise assumed to be independent and identically distributed (i.i.d.) with normal distribution:

$$x_{1i} = \underline{x_{1i}} + e_{1x}^{1i}, \quad y_{1i} = \underline{y_{1i}} + e_{iy}^{1i}, \quad x_{2i} = \underline{x_{2i}} + e_{1x}^{2i}, \quad \text{and} \quad y_{2i} = \underline{y_{2i}} + e_{iy}^{2i}, \quad (6)$$

where $e_{1x}^{1i}, e_{iy}^{1i}, e_{1x}^{2i}$ and $e_{iy}^{2i} \sim N(0, \sigma_n^2)$ and $\sigma_n$ is an unknown scale of measurement noise. The underlined variables denote the actual or noise-free locations of points in image plane.

We try to segment matching points undergoing $T$ from the static background in two images. Thus, the points undergoing $T$ are the inliers of the motion segmentation while the static points are the outliers. The fundamental matrix of points undergoing $T$ is computed using $[6, 16, 17]$:

$$F = A^{-\top} [T]_x R A^{-1}, \quad (7)$$

where $A$ is the camera calibration matrix, $R$ is the rotation matrix and $[T]_x$ is the skew-symmetric matrix of $T$ $[7-9]$. In the case of $T = [T_x, T_y, 0]^{\top}$, $R = I_3$ (representing zero rotation) and the focal length of the camera matrix is $f$ (equal focal length in $X$ and $Y$ direction is used and the
offset distance \( P_x \) and \( P_y \) are set to zero for simplicity), equation (7) yields:

\[
F_T = \frac{1}{f} \begin{pmatrix}
0 & 0 & T_y \\
0 & 0 & -T_x \\
-T_y & T_x & 0
\end{pmatrix}.
\] (8)

We assume that, a perfect estimator provides the true fundamental matrix given in (8). If \( F_T \) is known, the Sampson distances can be computed using (4). Substitution of real plus noise forms in (6) and \( F_T \) in (8) yields:

\[
d_i = \frac{T_y(\vec{x}_{2i} + e_{ix}^2 - \vec{x}_{1i} - e_{ix}^1) + T_x(y_{1i} + e_{iy}^1 - y_{2i} - e_{iy}^2)}{\sqrt{2(T_y^2 + T_x^2)}}.
\] (9)

For points undergoing \( T \) (\( i = 0, 1 \ldots n_i \)), the above expression without noise terms equals zero (according to (1) and because the true \( F_T \) is used to compute \( d_i \)). Thus, equation (9) can be simplified to:

\[
d_i = \frac{T_y(e_{ix}^2 - e_{ix}^1) + T_x(e_{iy}^1 - e_{iy}^2)}{\sqrt{2(T_y^2 + T_x^2)}}
\] (10)

The above distances turn out to be a linear combination of the i.i.d. noise samples. Therefore, they are also normally distributed with zero mean and variance \( \sigma_n^2 \) as the numerator and denominator cancel each other. For static background points (outliers) we have:

\[
\vec{x}_{1i} = \vec{x}_{2i} \quad \text{and} \quad y_{1i} = y_{2i}, \quad \text{for} \ i = n_i + 1, n_i + 2 \ldots n.
\] (11)

Calculation of the distances of the static points with respect to \( F_T \) using (9) and (11) results in the same formula as given in (10). Thus, the distances of the static background are also distributed according to \( N(0, \sigma_n^2) \). More precisely the distributions of \( d_i \)'s for points undergoing \( T \) and static background will be exactly the same.

The success of motion segmentation is determined by looking at the relative sizes of inlier and outlier distances. For the inliers (points undergoing \( T \)) to be successfully segmented, the smallest outlier (background point) distances should be sufficiently larger that the biggest inlier distances. Since the distributions of \( d_i \)'s for points undergoing \( T \) and static background are exactly the same, points undergoing \( T \) cannot be segmented from static background (if the
fundamental matrix motion model is used).

The segmentation of translation $T$ with $T_z \neq 0$ form static background is too complex to be derived theoretically. However, the results of our Monte Carlo experiments (presented in Section 4) verify that the Sampson distances of the points undergoing $T$ with $T_z \neq 0$ and static background are also normally distributed with zero mean and similar variances. Therefore they are also not separable from static background.

4 Conditions for motion-background segmentation

The non-separability of pure translations from its background (when using the fundamental matrix as motion model) implies that the separability of a motion from its background depends on its rotational part. We therefore aim to determine the required condition in term of minimum rotation angle for successful motion-background segmentation using Monte Carlo experiments. The correctness of these conditions was verified by studying the variance of the results of the experiments. The Monte Carlo experiments have two main objectives: verification of the non-separability of a pure translation from its static background, and determining the minimum rotation angle for successful segmentation. The Monte Carlo experiments were conducted with the assumption of having the accurate fundamental matrix (this is the same assumption in our analysis in Section.3). In addition, only the correct matches belonging either to the background or to the motion were assumed to exist in the data. In reality, the distances are distorted as they were calculated using the estimate of the fundamental matrix and there are mismatches due to the matching errors. Thus, the theoretical conditions for successful motion-background segmentation determined in this paper are the necessary conditions and not the sufficient conditions. For example, if in theory, a minimum rotation of $5^\circ$ is required for correct motion-background segmentation, in practice (due to imperfections such as the estimation error of the fundamental matrix and the presence of mismatches) rotations of more than $5^\circ$ will be required to guarantee a successful segmentation of the motion from the background. The analysis of the extent to which such imperfections affect the theoretical results is out of the scope of this paper and will be considered in future works.

In each experiment, the $X$, $Y$ and $Z$ coordinate of 2000 moving points $M_i$'s (the inlier part in the experiments) were randomly generated. These points were visible to a synthetic camera
with the following parameters:

\[
A = \begin{pmatrix}
703 & 0 & 256 \\
0 & 703 & 256 \\
0 & 0 & 1
\end{pmatrix}.
\] (12)

The projection of a point \(M_i\) in the first camera position was \(\mathbf{m}_{1i} = [I|0]M_i\) and \(\mathbf{m}_{2i} = [R|T]M_i\) in the second camera position after the camera makes a rotation \(R\) and translation \(T\). The rotation was assumed only around Z-axis denoted by \(\theta_z\). Based on the principle of duality, this scenario was equivalent to having a static camera and 2000 points moved according to rotation \(R\) and translation \(T\).

The static points representing the background (the outlier part in the experiments) were randomly added to both images based on the magnitude of the intended inlier ratio \(\epsilon\). Then all generated image points were perturbed with Gaussian noise of \(N(0,\sigma_n)\), where \(\sigma_n^2 = 1\). The true \(F\) of the points undergoing motion was calculated using equation (7) (by substitution of the known object motion \(R\) and \(T\) and camera matrix in (12)). Then distances \(d_i\)'s for all points (both moving and static points) were computed using the true \(F\) and \(d_i^2\)'s were used as residuals for segmentation using MSSE [21]. In MSSE, \(d_i^2\)'s were sorted in an ascending order and the scale estimate given by the smallest \(k_{th}\) distances is calculated using [21]:

\[
\sigma_k^2 = \frac{\sum_{i=1}^{k} d_i^2}{k - 1}.
\] (13)

While incrementing \(k\), \(d_{k+1}\) was detected as the distance of the first outlier if it is larger than 2.5 times the scale estimate given by the smallest \(k_{th}\) distances:

\[
d_{k+1}^2 > 2.5^2 \sigma_k^2.
\] (14)

With the above threshold, 99.4% of the moving points will be segmented if they are normally

\footnote{The derivation of similar results for rotations around \(X\) and \(Y\) axes would be straightforward as it follows similar steps.}
Repeat (inlier ratio $\epsilon = 30\%$ to $80\%$), (noise level $\sigma_n = 0.25$ to $2$) and (rotation angle $\theta_z = 0^\circ$ to $70^\circ$)

i. Generate 1000 random translations between $\pm 0.1m$

Repeat (translation $j = 1$ to 1000)
1. Generate 2000 random pairs of moving points according to translation $j$ and $\theta_z$
2. Generate random pairs of static points based on $\epsilon$
3. Perturb all points with Gaussian noise $N(0,\sigma_n^2)$
4. Calculate the true $F$ of the moving points
5. Calculate the Sampson distances using the true $F$
6. Sort the Sampson distances and perform segmentation using MSSE
7. Record the scale estimate given by the actual moving points (inlier scale) and by all image points (total scale) separately
8. Record the ratio of the number of segmented moving points to the actual number of moving points $\zeta$

End

ii. Calculate and record the mean and standard deviation of the 1000 inlier scales, total scales and $\zeta$’s

End

Figure 2: Pseudo code of the Monte Carlo experiments.

distributed [21]. In our experiments we iteratively compute the segmentation variable:

$$\rho_k = \frac{d_{k+1}^2}{2.5^2\sigma_k^2},$$  

and segment the moving points when $\rho_k \geq 1$.

After each segmentation, the ratio of the number of segmented moving points to the number of actual moving points, denoted by $\zeta$, was calculated and recorded. Every set of the experiments were repeated for 1000 times where translations in $x$, $y$ and $z$ directions were randomly selected and varied between -0.1 and 0.1 meter (equivalent to about $\pm 20$ pixels in the image plane). The pseudo code of the Monte Carlo experiments is given in Fig.2.

The first set of the experiments were designed to verify our earlier statement on non-separability of pure translational motions from static background. The experiments were conducted with random 2D translations ($T_z = 0$) in Case-1 and random 3D translations ($T_z \neq 0$) in Case-2 with $\epsilon = 80\%$, $\theta_z = 0$ and $\sigma_n = 1$ using 2000 randomly selected image points undergoing
Table 1: Inlier and total scale for pure translational motion for $\epsilon = 80\%$ and $\sigma_n = 1$.

<table>
<thead>
<tr>
<th>Translation magnitude, $[T_x, T_y, T_z]^\top$ cm</th>
<th>Inlier scale</th>
<th>Total scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case-1 [-9.79,-2.67,0]</td>
<td>1.0066</td>
<td>1.0082</td>
</tr>
<tr>
<td>Case-2 [-5.65,-2.36,-9.16]</td>
<td>1.0012</td>
<td>1.0075</td>
</tr>
</tbody>
</table>

Translation $T (\theta_z = 0)$. Two types of scale estimates were calculated: the scale estimate given by the actual moving points (called inlier scale) and the scale estimate given by all data points (called total scale). For two instances of the data samples generated in Case-1 and Case-2, we have plotted the histogram of $d_i$’s for all image points (moving and static points) and the segmentation variable $\rho_k$ in Fig.3 and 4, respectively. The random translations applied in these instances and the inlier and total scales are given in Table 1 to verify the similarity of $d_i$’s for the moving points and all image points.

It was observed that the inlier and total scales were very close to 1, reminding that the true corresponding point locations were perturbed by the Gaussian noise of $N(0,\sigma_n^2)$ (where $\sigma_n = 1$) in the experiments. Similar magnitudes of inlier scales and total scales for 2D and 3D translations suggests that pure translational motions are not separable from static (or nearly static) background. This condition is also demonstrated by the close resemblance of the distribution of $d_i$’s for all image points (moving and static points) in both cases to $N(0,1)$ as shown in Fig.3(a) and 3(b). The non-separability is also shown by the graphs of the segmentation variable $\rho_k$ which crosses the line 1 at about 2450 points (the total number of points is 2500) in Fig.4(a) and 4(b) instead of at 2000 for correct segmentation.

The second set of the experiments were designed to study the relationship between the separability of pure translations from static background and the inlier ratio $\epsilon$. Maintaining 2000 randomly selected image points undergoing pure translation ($\theta_z = 0$) and $\sigma_n = 1$, the experiments were repeated by varying $\epsilon$ from 20% to 80% (1000 randomly selected 3D translations were used in each value of $\epsilon$). The mean and standard deviation of the inlier and total scales are summarised in Table 2. The readings of the mean and standard deviation of the inlier and total scales were nearly equal to 1 and 0, respectively, for various $\epsilon$ values, showing that the inlier and the total scales were almost consistent for all 1000 sets of random translations with each $\epsilon$. Therefore, the non separability of pure translations from static background analysis in Section 4 was verified for all $\epsilon$’s examined in these experiments.
Figure 3: Distribution of Sampson distances $d_i$’s for all image points (moving and static).

Figure 4: The segmentation variables $\rho_k$.

Table 2: True scale and total scale of pure translational motions for various inlier ratio $\epsilon$.

<table>
<thead>
<tr>
<th>$\epsilon$, %</th>
<th>No of outliers</th>
<th>Mean of inlier scale</th>
<th>$\sigma$ of inlier scale</th>
<th>Mean of total scale</th>
<th>$\sigma$ of total scale</th>
</tr>
</thead>
<tbody>
<tr>
<td>80</td>
<td>500</td>
<td>1.0001</td>
<td>0.0156</td>
<td>0.9999</td>
<td>0.0139</td>
</tr>
<tr>
<td>70</td>
<td>857</td>
<td>1.0003</td>
<td>0.0161</td>
<td>1.0001</td>
<td>0.0132</td>
</tr>
<tr>
<td>60</td>
<td>1333</td>
<td>0.9994</td>
<td>0.0160</td>
<td>0.9995</td>
<td>0.0124</td>
</tr>
<tr>
<td>50</td>
<td>2000</td>
<td>0.9994</td>
<td>0.0154</td>
<td>0.9997</td>
<td>0.0113</td>
</tr>
<tr>
<td>40</td>
<td>3000</td>
<td>0.9998</td>
<td>0.0159</td>
<td>1.0000</td>
<td>0.0104</td>
</tr>
<tr>
<td>30</td>
<td>4667</td>
<td>1.0009</td>
<td>0.0164</td>
<td>1.0002</td>
<td>0.0093</td>
</tr>
<tr>
<td>20</td>
<td>8000</td>
<td>0.9994</td>
<td>0.0161</td>
<td>0.9996</td>
<td>0.0070</td>
</tr>
</tbody>
</table>

The third set of the experiments were designed to determine the separability conditions for a motion from its static (or nearly static) background. Each experiment involved a motion including a rotation $\theta_z$ around the Z axis (ranging from 0 to $70^\circ$) followed by a random 3D translation, for various inlier ratios $\epsilon$ (30% to 80%) and noise levels $\sigma_n$ (0.25 to 2). The number of moving points was maintained at 2000 image points (which were randomly selected), and
1000 set of random translations were examined in each experiment for each $\theta_z$, $\epsilon$ and $\sigma_n$. In order to validate the independence of the results from the camera matrix $A$, in some cases, the camera matrix in equation (12) was varied both homogenously (keeping the focal lengths in $X$ and $Y$ directions equal) and non-homogenously. In these experiments, we examined the following variations of the camera matrix:

$$A_1 = \begin{pmatrix} 492.1 & 0 & 256 \\ 0 & 492.1 & 256 \\ 0 & 0 & 1 \end{pmatrix} \quad \text{and} \quad A_2 = \begin{pmatrix} 527.3 & 0 & 256 \\ 0 & 597.6 & 256 \\ 0 & 0 & 1 \end{pmatrix}. \quad (16)$$

Fig.5 and 6 show the mean and standard deviation of $\zeta$’s (ratio of the number of segmented moving points to the number of actual moving points) versus $\theta_z$ at $\sigma_n=1$ for $\epsilon=80\%$ and $40\%$ for different camera matrixes. For perfect segmentation $\zeta$ should be equal to one, i.e. the number of segmented moving points is the same as the number of actual moving points.

The very small standard deviations of $\zeta$’s in Fig.5 and 6 indicated that the segmentation results were independent of the locations of moving points in the image plane since these points (2000 pairs of point) were randomly selected in each of the 1000 iterations. In addition, the very small standard deviations of $\zeta$’s also showed that the consistency of the segmentation results which is coherent with our earlier statement that the success of motion-background segmentation depends only on the rotational part and independent of the translational part of the motion (the translational part of the motion were randomly selected in each of the 1000 iterations). Furthermore, the close resemblance of the $\zeta$-$\theta_z$ plots shown in Fig.5(a)-6(a) and in Fig.5(b)-6(b) (in terms of both the magnitude of the mean and the standard deviation of $\zeta$ values) confirmed that the separability of a general motion from its background was also independent of the camera matrix.

It was also observed in Fig.5 and 6 that for small rotations, some background points were mixed with the moving points and segmented as part of the moving points ($\zeta > 1$). In such cases, the inaccurate dichotomy between moving and background points would result in an incorrect motion estimate. We have examined two thresholds for the segmentation ratio $\zeta$ to be considered correct: $\zeta_1 = 1.05$ and $\zeta_2 = 1.10$. Using these thresholds mean that it was accepted for 5% and 10% of background points with small distances to be segmented as moving points. For each
threshold, we record the minimum rotation angle required for successful motion-background segmentation (denoted by $\tilde{\theta}_z$). Those angles for various inliers ratio and noise levels are plotted in Fig.7 and 8.

The results also showed that the minimum rotation angle $\tilde{\theta}_z$ required for segmentation increased for smaller inlier ratios and/or larger noise levels. More precisely, the separability of a general motion from its background became more difficult when more points in the background
and/or higher level of noise were involved. This is because, as the contamination of points belonging to the background in the image increased ($\epsilon$ decreased), the density of background points residuals became more spread. Thus, the likelihood of background points with small residuals being segmented as moving points increased. Furthermore, when the noise levels were larger, the maximum distance of moving points became larger resulting in more background points with distances smaller than the maximum distance of moving points. Therefore more background points were likely to be segmented as moving points. Hence, larger magnitude of $\tilde{\theta}_z$ was required to produce sufficiently distinct moving points distances and background points distances for correct segmentation.

5 Experiment using real images

In a controlled experiment using real images we examined the validity of the Monte Carlo experiments results on the non-separability of pure translations from static background (when fundamental matrix is used) and our statement on the success of motion-background segmentation depends on the magnitude of $\theta_z$. For the validation of theoretical and Monte Carlo experiments results (in Section 3 and 4), we need to eliminate the effect of incorrect estimation of fundamental matrix, presence of incorrect SIFT matches and image distortions on the segmentation results. Therefore, we have used the actual fundamental matrix of the motion involved in the experiments and manually eliminated the incorrect SIFT matches. In addition the camera was also calibrated to determine the parameters for calculating the actual fundamental matrix, reduce image distortions and generate accurate motions using the camera’s principle point as
the reference point. The experiments consist of three steps:

1. Camera calibration and image acquisition: Firstly, a camera was calibrated using a publicly available camera calibration toolbox by Bouguet [26]. Then a 3D object was moved in incremental rotations $\theta_z$ around Z axis from $0^\circ$ to $10^\circ$ ($2^\circ$ in each increment) each followed by 20 sets of distinct 3D translations. The principles point of the camera from camera calibration was used as the reference point for the object’s motions. Images of the object and its static background were taken before and after each motion. In total 120 pairs of images were used in the experiments.

2. Image data preparation: The image distortions were reduced from all images using the distortion parameter from camera calibration. Then, the corresponding image points belonging to the object in motion and its background were extracted from each pair of images using an implementation of SIFT algorithm by Lowe [1, 2]. After that, all incorrect matches were manually eliminated. The inlier ratio in each pair of images were varied ($\epsilon$ from 35% to 80%) by removing some of the background points while maintaining the moving points.

3. Segmentation analysis: The actual $F$ for the points belonging to the moving object for each pair of images was calculated using (7) with known object’s motion and the camera matrix based on camera calibration. Then the distances $d_i$’s were calculated for all image points using (4) and the true $F$. Motion segmentation was performed to identify the moving points in each pair of images using MSSE (14) with $d_i^2$ as the residuals. The noise level $\sigma_n$ in each pair of images was estimated using equation (13) and $d_i$’s of the actual moving points. Finally, the ratio of the segmented over actual moving points $\zeta$ was calculated and recorded for each pair of images for all $\theta_z$ (from $0^\circ$ to $10^\circ$) and $\epsilon$ (from 35% to 80%).

Sample result of the real image experiments are shown in Fig.9 where the object was moved according to $\theta_z = 0^\circ$, $4^\circ$ and $8^\circ$ followed by a constant 3D translation of $T = [-59-82-39]^\top$ mm. The inlier ratio (ratio of moving points over all image points) was set to 35%. It was observed that almost all image points were selected as moving points when the motion was a pure translation as shown in Fig.9(d) and the value $\zeta = 2.40$ ($\zeta = 1$ for perfect segmentation). In addition, the distribution of $d_i$’s of moving points and static background were well mixed and cannot be
Figure 9: Corresponding image points of image 1 and image 2 superimposed on image 1 for \( \theta_z = 0^\circ, 4^\circ \) and \( 8^\circ \) followed by \( T = [−59 \ -82 \ -39]^T \) mm with \( \epsilon = 35\% \) in (a), (b) and (c). The segmented image points belonging to the motion in (d), (e) and (f) with \( \zeta = 2.40, 1.12 \) and \( 1.03 \) respectively. Histogram of \( d_i \)'s for all image points in (g), (h) and (i).

distinguished from each other (as shown in Fig.9(g)). As \( \theta_z \) increased to \( 4^\circ \) and \( 8^\circ \) while \( T \) remains constant, \( \zeta \) reduced to 1.12 and 1.03 and the moving points were correctly segmented as shown in Fig.9(e) and 9(f). Successful segmentation of the moving points became possible because \( d_i \)'s of static points were more spread and can be distinguished from \( d_i \)'s of moving points in Fig.9(h) and 9(i). The mean and standard deviation of \( \zeta \) versus \( \theta_z \) (20 \( \zeta \) for each \( \theta_z \)) are presented in Fig.10. It could be observed that, the segmentation was inaccurate when \( \theta_z = 0^\circ \) (\( \zeta > 1.10 \)) and as \( \theta_z \) increased the ratio \( \zeta \) reduced and approaching \( \zeta = 1 \) indicating correct segmentation. These experimental results and observations verified the non-separability of pure translations from static background and showed that the success of motion-background segmentation depends on the magnitude of the rotation angle \( \theta_z \).
Figure 10: Mean and standard deviation of $\zeta$ vs $\theta_z$ for various $\epsilon$.

![Graphs showing mean and standard deviation for different $\epsilon$ values.](image)

Figure 11: $\tilde{\theta}_z$ vs $\epsilon$ for Monte Carlo and real image experiments.

![Graphs showing $\tilde{\theta}_z$ vs $\epsilon$ for different $\zeta$ values.](image)

The minimum rotation angle $\tilde{\theta}_z$ required for correct segmentation (corresponding to $\zeta_1 = 1.05$ and $\zeta_2 = 1.10$) have been extracted from the segmentation results shown in Fig.10 for all inlier ratios. The average noise level ($\sigma_n$) of all image pairs (estimated from equation (13) and the actual $d_i$’s of the moving points) was approximately $\sigma_n \approx 0.39$. The extracted $\tilde{\theta}_z$’s from real image experiments were compared with $\tilde{\theta}_z$’s from Monte Carlo experiment (from Fig.7) for noise levels $\sigma_n = 0.25$ and $\sigma_n = 0.5$ (for both to $\zeta_1$ and $\zeta_2$) in Fig.11. It was observed that $\tilde{\theta}_z$’s increased as $\epsilon$ decreased for both $\zeta_1$ and $\zeta_2$. This indicates that the motion background segmentation became more difficult when more static points were involved. The similar trend of $\tilde{\theta}_z$ vs $\epsilon$ (in the experiment using real image to the ones in Monte Carlo experiments) verifies
the Monte Carlo experiments results summarised in the graphs shown in Fig.7 and 8.

6 Conclusions

In this paper, it was shown that a pure translational motion is not separable from static background in motion segmentation using fundamental matrix. The success of motion-background segmentation depends on the rotation angle of the motion. This relationship was explored both by theoretical analysis and experimentation with synthetic and real images. Monte Carlo experiments using synthetic images provided the minimum required rotation angle for correct motion-background segmentation. It was also shown that the minimum required rotation angles increased as the inlier ratio decreased and the noise level increased. This condition occurs because when more background points involved or the noise level increased, the likelihood for background points to be wrongly segmented as moving points increased. Therefore resulting in a higher magnitude of minimum required rotation angle for correct motion-background segmentation. Based on the experimental results using both synthetic and real images, the fundamental matrix motion model are not suitable for motion-background segmentation when the rotation angle of the motion is less than the required angle for correct segmentation.

References


