Optimal Segmentation of Dynamic Scenes from Two Perspective Views*

René Vidal Shankar Sastry
Department of EECS, University of California, Berkeley, CA, 94720

Abstract

We present a novel algorithm for optimally segmenting dynamic scenes containing multiple rigidly moving objects. We cast the motion segmentation problem as a constrained nonlinear least squares problem which minimizes the reprojection error subject to all multibody epipolar constraints. By converting this constrained problem into an unconstrained one, we obtain an objective function that depends on the motion parameters only (fundamental matrices), but is independent on the segmentation of the image features. Therefore, our algorithm does not iterate between feature segmentation and single body motion estimation. Instead, it uses standard nonlinear optimization techniques to simultaneously recover all the fundamental matrices, without prior segmentation. We test our approach on a real sequence.

1. Introduction

Segmentation of dynamic scenes refers to the problem of simultaneously estimating the motion of multiple rigidly moving objects from the measurements collected by a static or moving camera. This is a challenging problem in visual motion analysis, because it requires the simultaneous estimation of an unknown number of motion models, without knowing which measurements correspond to which model.

Prior work on motion segmentation subdivides the problem in two stages: *feature segmentation* and *single body motion estimation*. In the first stage, image measurements corresponding to the same motion model are grouped together using various clustering algorithms, e.g., K-means. In the second stage, a different motion model is estimated from the measurements corresponding to each group.

This two-stage approach to motion segmentation is clearly not optimal in the presence of noise. In order to obtain an optimal segmentation, probabilistic approaches model the scene as a mixture of motion models in which the measurements are corrupted with noise, e.g., zero-mean and Gaussian. The membership of the data is also modeled with a probability distribution by using the so-called mixing proportions. Unfortunately, the simultaneous maximum likelihood estimation of both mixture and motion parameters is in general a hard problem. Therefore, most of the existing methods solve the problem in two stages. One first computes the expected value of the mixing proportions given a prior on the motion parameters and then maximizes the likelihood of the motion parameters given a prior on the grouping of the data. This is usually done in an iterative manner using the Expectation Maximization (EM) algorithm.

One of the main reasons for using either of these two-stage approaches (iterative or not) is that the problem of estimating a single motion model from image measurements is reasonably well understood, both from a geometric and from an optimization point of view. For example, it is well known that two views of a scene are related by the so-called epipolar constraint and that multiple views are related by the so-called multilinear constraints. These constraints can be naturally used to estimate the motion parameters using linear techniques, such as the eight-point algorithm and its generalizations. In the presence of noise, many optimal nonlinear algorithms for single body motion estimation have been proposed. For example, see [8, 12, 14, 18] for the discrete case and [10] for the differential case.

The two-view geometry of multiple moving objects is, on the other hand, not as well understood. In fact, the first generalizations of the eight-point algorithm to multiple motions were not known until very recently [20, 16]. The work of [16] showed that it is possible to simultaneously recover *multiple* motion models, *without* previously segmenting the image measurements. The key is to use a geometric constraint that is satisfied by all the measurements, regardless of the group to which they belong, the so-called *multibody epipolar constraint*. The number of moving objects n can be obtained from a rank constraint on the measurements, and multiple fundamental matrices can be simultaneously recovered using a novel polynomial factorization technique. The solution can be computed in polynomial time, and is closed form if and only if $n \leq 4$. Similar techniques can be used for direct motion segmentation from image intensities by using the so-called *multibody affine constraint* [17].

Unfortunately, these algebraic geometric techniques are more sensitive to noise than the eight-point algorithm, because in order to obtain a linear solution they use an overparameterized representation of the motion parameters via the so-called *multibody fundamental matrix*. Although one could use a minimal representation by minimizing the error defined by the algebraic constraints, we will show in this paper that this is not optimal in the presence of noise.

*We thank Drs. Y. Ma and C. Geyer for their comments. Work funded by grants ONR N00014-00-1-0621 and DARPA F33615-98-C-3614.

1.1. Contributions of this paper

In this paper, we present a novel approach to 3-D motion segmentation that is *optimal* in the presence of noise and does *not* iterate between feature segmentation and single body motion estimation. In Section 3 we show that one can eliminate the feature segmentation stage by using the multibody epipolar constraint, which is by definition independent on the segmentation of the image measurements. In Section 4 we show that using this algebraic constraint as the *objective* function is not optimal. Therefore, we cast the motion segmentation problem as a constrained nonlinear least squares problem which minimizes the reprojection error subject to all multibody epipolar constraints. By converting this constrained problem into an unconstrained one, we obtain an objective function that depends on the motion parameters only (fundamental matrices) and is independent on the segmentation of the image data. We then use standard nonlinear optimization techniques to simultaneously recover all the fundamental matrices, without prior feature segmentation. In Section 5 we evaluate the performance of the algorithm with respect to the number of motions and the amount of noise. We also test the proposed approach by segmenting a real sequence. Section 6 concludes the paper.

1.2. Previous work

Geometric approaches to 3-D motion estimation and segmentation based on 2-D imagery include: multiple points moving linearly with constant speed [4, 9] or in a pencil of planes [11], multiple moving objects seen by an orthographic camera [1, 6], self-calibration from multiple motions [3, 5], reconstruction of multiple translating planes [19], and segmentation of two rigid motions from two perspective views [20].

Probabilistic approaches to 3-D motion segmentation model the scene as a mixture of probabilistic models. The number of models is estimated using model selection techniques [13, 7], and the motion models are estimated using an iterative process that alternates between segmentation and motion estimates [13, 2] using, e.g., the EM algorithm.

2. Multibody Structure from Motion

We consider two images of a scene containing an *unknown* number n of independently and rigidly moving objects. We describe the motion of each object relative to the camera between the two frames with the rank-2 fundamental matrix $F_i \in \mathbb{R}^{3\times 3}$ associated with the motion of object $i = 1, \ldots, n$. We assume that the motions of the objects are such that all the fundamental matrices are distinct and different from zero, and hence the relative translation between the two image frames is non-zero.

The projection of a point $q^j \in \mathbb{R}^3$ onto the image frame I_k is denoted as $\boldsymbol{x}_k^j \in \mathbb{P}^2$, for $j = 1, \ldots, N$ and $k = 1, 2$. In order to avoid degenerate cases, we will assume that the image points $\{\boldsymbol{x}_k^j\}$ correspond to 3-D points $\{q^j\}$ in general configuration in $\mathbb{R}^3$, i.e. they do not all lie in any critical surface, for example. We will drop the superscript when we refer to a generic image pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$. Also, we will always use the homogeneous representation $\boldsymbol{x} = [x, y, z]^T \in \mathbb{R}^3$ to refer to an arbitrary image point in $\mathbb{P}^2$.

We define the *multibody structure from motion problem* as follows:

Problem 1 (Multibody structure from motion problem) *Given a set of image pairs $\{(\boldsymbol{x}_1^j, \boldsymbol{x}_2^j)\}_{j=1}^N$ corresponding to an unknown number of independently and rigidly moving objects that satisfy the assumptions above, estimate the number of independent motions n, the fundamental matrices $\{F_i\}_{i=1}^n$, and the segmentation of the image pairs, i.e. the object to which each image pair belongs.*

3. Multibody Epipolar Geometry

In this section, we describe the two-view geometry of multiple moving objects. We introduce the multibody epipolar constraint and the multibody fundamental matrix, and show how they can be used to estimate the number of independent motions and the individual fundamental matrices in the absence of noise in the image measurements.

3.1. The multibody epipolar constraint

Let $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ be an arbitrary image pair corresponding to any motion. Then, there exists a fundamental matrix F_i such that the epipolar constraint $\boldsymbol{x}_2^T F_i \boldsymbol{x}_1 = 0$ holds. Thus, regardless of the object to which the image pair belongs, the following constraint must be satisfied by the number of independent motions n, the relative motions of the objects $\{F_i\}_{i=1}^n$ and the image pair $(\boldsymbol{x}_1, \boldsymbol{x}_2)$

$$\boxed{\mathcal{E}(\boldsymbol{x}_1, \boldsymbol{x}_2) \doteq \prod_{i=1}^{n} \left(\boldsymbol{x}_2^T F_i \boldsymbol{x}_1\right) = 0.} \tag{1}$$

We call this constraint the *multibody epipolar constraint*, since it is a natural generalization of the epipolar constraint valid for $n = 1$ to the case of multiple motions.

3.2. The multibody fundamental matrix

The multibody epipolar constraint converts Problem 1 into one of solving for the number of independent motions n and the fundamental matrices $\{F_i\}_{i=1}^n$ from the *nonlinear* equation (1). This nonlinear constraint defines a homogeneous polynomial of degree n in either $\boldsymbol{x}_1$ or $\boldsymbol{x}_2$. For example, if we let $\boldsymbol{x}_1 = [x_1, y_1, z_1]^T$, then equation (1) viewed as a function of $\boldsymbol{x}_1$ can be written as a linear combination of the

following monomials $\{x_1^n, x_1^{n-1}y_1, x_1^{n-1}z_1, \ldots, z_1^n\}$. It is readily seen that there are a total of

$$M_n \doteq (n+1)(n+2)/2 \tag{2}$$

different monomials. Thus, if we use the Veronese map of degree n, $\nu_n : \mathbb{P}^2 \rightarrow \mathbb{P}^{M_n-1}$, to map $[x_1, y_1, z_1]^T$ to all its monomials of degree n $[x_1^n, x_1^{n-1}y_1, x_1^{n-1}z_1, \ldots, z_1^n]^T$, then we can rewrite the multibody epipolar constraint (1) in bilinear form as (see [16] for the proof)

$$\boxed{\nu_n(\boldsymbol{x}_2)^T \mathcal{F} \nu_n(\boldsymbol{x}_1) = 0,} \tag{3}$$

where $\mathcal{F} \in \mathbb{R}^{M_n \times M_n}$ is a matrix representation of the symmetric tensor product of all the fundamental matrices $\{F_i\}_{i=1}^n$. We call the matrix $\mathcal{F}$ the *multibody fundamental matrix* since it is a natural generalization of the fundamental matrix to the case of multiple moving objects. Since equation (3) resembles the bilinear form of the epipolar constraint for a single rigid body motion, we will refer to both equations (1) and (3) as the *multibody epipolar constraint*.

3.3. Estimating the number of motions n and the multibody fundamental matrix $\mathcal{F}$

Since the multibody epipolar constraint (3) is *linear* in $\mathcal{F}$, we can rewrite it as $(\nu_n(\boldsymbol{x}_2) \otimes \nu_n(\boldsymbol{x}_1))^T \boldsymbol{f} = 0$, where $\boldsymbol{f} \in \mathbb{R}^{M_n^2}$ is the stack of the columns of $\mathcal{F}$ and $\otimes$ represents the Kronecker product. Therefore, given a collection of image pairs $\{(\boldsymbol{x}_1^j, \boldsymbol{x}_2^j)\}_{j=1}^N$, the vector $\boldsymbol{f}$ satisfies

$$\boxed{L_n \boldsymbol{f} = 0,} \tag{4}$$

where the j^{th} row of $L_n \in \mathbb{R}^{N \times M_n^2}$ equals $(\nu_n(\boldsymbol{x}_2^j) \otimes \nu_n(\boldsymbol{x}_1^j))^T$, for $j = 1, \ldots, N$. In order to determine $\boldsymbol{f}$ uniquely (up to a scale factor) from (4), we must have that

$$\text{rank}(L_n) = M_n^2 - 1. \tag{5}$$

This rank constraint on L_n provides an effective criterion to determine the number of independent motions n from the given image pairs. Let $L_i \in \mathbb{R}^{N \times M_i^2}$ be the matrix in (4), but computed with the Veronese map ν_i of degree $i \geq 1$. We showed in [16] that if $N \geq M_n^2 - 1$ points in general configuration in 3-D are given and at least 8 points correspond to each motion, then $\text{rank}(L_i) = M_i$ if $i < n$, $\text{rank}(L_i) = M_i - 1$ if $i = n$ and $\text{rank}(L_i) < M_i - 1$ if $i > n$. Therefore, the number of independent motions n is given by

$$\boxed{n \doteq \min\{i : \textit{rank}(L_i) = M_i^2 - 1\}.} \tag{6}$$

Given n, we can linearly solve for the multibody fundamental matrix $\mathcal{F}$ from (4). Notice that the minimum number of image pairs needed is $N \geq M_n^2 - 1$, which grows in the order of $O(n^4)$ for large n. However, there are only $O(n)$ unknowns in the n fundamental matrices $\{F_i\}_{i=1}^n$.

3.4. Estimating the fundamental matrices

Given the multibody fundamental matrix $\mathcal{F}$ and the number of independent motions n, the rest of the problem is to recover the motion parameters (or fundamental matrices) and the segmentation of the image points. Mathematically, this is equivalent to factoring the multibody epipolar constraint into the product of n bilinear forms, i.e.

$$\nu_n(\boldsymbol{x}_2)^T \mathcal{F} \nu_n(\boldsymbol{x}_1) = \prod_{i=1}^{n} (\boldsymbol{x}_2^T F_i \boldsymbol{x}_1). \tag{7}$$

In [16] we showed how to solve this problem from the epipoles of each fundamental matrix and the epipolar lines associated with each image point. The estimation of epipoles and epipolar lines is based on the factorization of a given homogeneous polynomial of degree n in 3 variables with real coefficients into n distinct polynomials of degree 1 also with real coefficients. We showed that such a problem can be solved in polynomial time using linear algebraic techniques. Once the epipoles and the epipolar lines have been estimated, the estimation of individual fundamental matrices becomes a simple *linear* problem from which the segmentation of the image points is automatically obtained.

Since this algorithm naturally generalizes the well-known eight-point algorithm to multiple moving objects, we will refer to it as the *multibody linear algorithm*.

Remark 1 (Pure translation case) *When all the objects undergo a purely translational motion, one can directly recover the translation of each object relative to the camera by applying the same polynomial factorization algorithm to the known epipolar lines, as described in [15].*

Remark 2 (Segmenting translational and affine motions) *A similar factorization technique can be used for segmenting a dynamic scene directly from image intensities rather than from feature points. The case of translational motions is equivalent to factoring the multibody constant brightness constraint [15] and the case of affine motions is equivalent to factoring the multibody affine constraint [17].*

4. Optimal 3-D Motion Segmentation

The multibody linear algorithm provides an algebraic geometric solution to the problem of estimating a collection of fundamental matrices $\{F_i\}_{i=1}^n$ from image pairs $\{(\boldsymbol{x}_1^j, \boldsymbol{x}_2^j)\}_{j=1}^N$. In essence, the algorithm solves the set of nonlinear equations $\prod_{i=1}^n (\boldsymbol{x}_2^{jT} F_i \boldsymbol{x}_1^j) = 0$, $j = 1, \ldots, N$, in a "linear" fashion by embedding the image pairs into a higher-dimensional space via the Veronese map.

However, the multibody linear algorithm provides a linear solution at the cost of neglecting the internal nonlinear structure of the multibody fundamental matrix $\mathcal{F}$. For example, the algorithm solves for $M_n^2 - 1$ unknowns in $\mathcal{F} \in \mathbb{R}^{M_n \times M_n}$ from equation (4), even though there are

only $8n$ unknowns in the fundamental matrices $\{F_i\}_{i=1}^n$ ($5n$ in the calibrated case). In practice, solving for an over-parameterized representation of the multibody fundamental matrix can be very sensitive in the presence of noise.

One way of resolving this problem is to replace the multibody linear algorithm by the nonlinear least squares problem

$$\min_{F_1,\ldots,F_n} \sum_{j=1}^{N} (\nu_n(\boldsymbol{x}_2^j)^T \mathcal{F} \nu_n(\boldsymbol{x}_1^j))^2 = \sum_{j=1}^{N} \prod_{i=1}^{n} (\boldsymbol{x}_2^{jT} F_i^T \boldsymbol{x}_1^j)^2. \quad (8)$$

Minimizing this algebraic error in fact provides a more robust estimate of the fundamental matrices, because it uses a minimal representation of the unknowns. However, the solution to this optimization problem is not optimal, because the algebraic error in (8) does not coincide with the negative log-likelihood of the data given the parameters.

In this section, we derive an optimal algorithm for estimating the fundamental matrices when the image pairs are corrupted with i.i.d. zero-mean Gaussian noise. We show that the optimal solution can be obtained by minimizing a function similar to the algebraic error in (8), but properly normalized. We cast the motion segmentation problem as a constrained nonlinear least squares problem which minimizes the reprojection error subject to all the multibody epipolar constraints. Since the multibody epipolar constraint is satisfied by *all* image pairs, irrespective of the segmentation, we do not need to model the membership of the image pairs with a probability distribution. Hence, we do not need to iterate between feature segmentation and single body motion estimation, as in EM-like techniques. In fact, the segmentation (E step) is *algebraically eliminated* by the multibody epipolar constraint, which leads to an objective function that depends only on the motion parameters.

Let $\{(\boldsymbol{x}_1^j, \boldsymbol{x}_2^j)\}_{j=1}^N$ be the given collection of noisy image pairs. We would like to find a collection of fundamental matrices $\{F_i\}_{i=1}^n$ such that the corresponding noise free image pairs $\{(\tilde{\boldsymbol{x}}_1^j, \tilde{\boldsymbol{x}}_2^j)\}_{j=1}^N$ satisfy the multibody epipolar constraint $\nu_n(\tilde{\boldsymbol{x}}_2^j)^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j) = 0$. Since for the Gaussian noise model the negative log-likelihood is equal to the reprojection error, we obtain the constrained optimization problem[1]

$$\begin{array}{ll} \min & \sum_{j=1}^N \|\tilde{\boldsymbol{x}}_1^j - \boldsymbol{x}_1^j\|^2 + \|\tilde{\boldsymbol{x}}_2^j - \boldsymbol{x}_2^j\|^2 \\ \text{subject to} & \nu_n(\tilde{\boldsymbol{x}}_2^j)^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j) = 0 \quad j = 1, \ldots, N. \end{array} \quad (9)$$

By using Lagrange multipliers $\lambda^j \in \mathbb{R}$ for each constraint, the above optimization problem is equivalent to minimizing

$$\sum_{j=1}^{N} \|\tilde{\boldsymbol{x}}_1^j - \boldsymbol{x}_1^j\|^2 + \|\tilde{\boldsymbol{x}}_2^j - \boldsymbol{x}_2^j\|^2 + \lambda^j \nu_n(\tilde{\boldsymbol{x}}_2^j)^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j). \quad (10)$$

[1]Notice that the optimization problem (9) does not include as an additional constraint the fact that the third entry of each image point $\boldsymbol{x} = [x, y, 1]^T \in \mathbb{R}^3$ is equal to one. We implicitly eliminate such constraints and their associated Lagrange multipliers by left-multiplying the partial derivatives of the Lagrangian (10) by the projection matrix Λ in (11).

Let

$$\Lambda = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} = [e_3]_\times^T [e_3]_\times, \quad (11)$$

with $e_3 = [0, 0, 1]^T \in \mathbb{R}^3$, be the projection matrix eliminating the third entry of any image point $\boldsymbol{x} = [x, y, 1]^T \in \mathbb{R}^3$. Since $\Lambda(\tilde{\boldsymbol{x}} - \boldsymbol{x}) = (\tilde{\boldsymbol{x}} - \boldsymbol{x})$, after multiplying the partial derivatives of the Lagrangian (10) with respect to $\tilde{\boldsymbol{x}}_1^j$ and $\tilde{\boldsymbol{x}}_2^j$ by the projection matrix Λ we obtain

$$2(\tilde{\boldsymbol{x}}_1^j - \boldsymbol{x}_1^j) + \lambda^j \Lambda \left(D\nu_n(\tilde{\boldsymbol{x}}_1^j) \right)^T \mathcal{F}^T \nu_n(\tilde{\boldsymbol{x}}_2^j) = 0, \quad (12)$$

$$2(\tilde{\boldsymbol{x}}_2^j - \boldsymbol{x}_2^j) + \lambda^j \Lambda \left(D\nu_n(\tilde{\boldsymbol{x}}_2^j) \right)^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j) = 0, \quad (13)$$

where $D\nu_n(\boldsymbol{x}) \in \mathbb{R}^{M_n \times 3}$ is the Jacobian of ν_n.

For ease of notation, let us also define

$$\boldsymbol{g}_1^j = (D\nu_n(\tilde{\boldsymbol{x}}_1^j))^T \mathcal{F}^T \nu_n(\tilde{\boldsymbol{x}}_2^j), \; \boldsymbol{g}_2^j = (D\nu_n(\tilde{\boldsymbol{x}}_2^j))^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j).$$

Then, since $\Lambda^T \Lambda = \Lambda^2 = \Lambda$, after left-multiplying (12) and (13) by $\boldsymbol{g}_1^{jT} \Lambda$ and $\boldsymbol{g}_2^{jT} \Lambda$, respectively, we obtain

$$2\boldsymbol{g}_1^{jT} \Lambda(\tilde{\boldsymbol{x}}_1^j - \boldsymbol{x}_1^j) + \lambda^j \|[e_3]_\times \boldsymbol{g}_1^j\|^2 = 0, \quad (14)$$

$$2\boldsymbol{g}_2^{jT} \Lambda(\tilde{\boldsymbol{x}}_2^j - \boldsymbol{x}_2^j) + \lambda^j \|[e_3]_\times \boldsymbol{g}_2^j\|^2 = 0. \quad (15)$$

Since $\Lambda(\tilde{\boldsymbol{x}} - \boldsymbol{x}) = (\tilde{\boldsymbol{x}} - \boldsymbol{x})$, $D\nu_n(\tilde{\boldsymbol{x}})\tilde{\boldsymbol{x}} = n\nu_n(\tilde{\boldsymbol{x}})$ and $\nu_n(\tilde{\boldsymbol{x}}_2)\mathcal{F}\nu_n(\tilde{\boldsymbol{x}}_1) = 0$, we obtain $\boldsymbol{g}_1^{jT} \tilde{\boldsymbol{x}}_1^j = \boldsymbol{g}_2^{jT} \tilde{\boldsymbol{x}}_2^j = 0$. Therefore, we have

$$-2\boldsymbol{g}_1^{jT} \boldsymbol{x}_1^j + \lambda^j \|[e_3]_\times \boldsymbol{g}_1^j\|^2 = 0, \quad (16)$$

$$-2\boldsymbol{g}_2^{jT} \boldsymbol{x}_2^j + \lambda^j \|[e_3]_\times \boldsymbol{g}_2^j\|^2 = 0, \quad (17)$$

from which we can solve for λ^j as

$$\frac{\lambda^j}{2} = \frac{\boldsymbol{g}_1^{jT} \boldsymbol{x}_1^j + \boldsymbol{g}_2^{jT} \boldsymbol{x}_2^j}{\|[e_3]_\times \boldsymbol{g}_1^j\|^2 + \|[e_3]_\times \boldsymbol{g}_2^j\|^2}. \quad (18)$$

Similarly, after left-multiplying (12) by $(\tilde{\boldsymbol{x}}_1^j - \boldsymbol{x}_1^j)^T$ and (13) by $(\tilde{\boldsymbol{x}}_2^j - \boldsymbol{x}_2^j)^T$ we get

$$2\|\tilde{\boldsymbol{x}}_1^j - \boldsymbol{x}_1^j\|^2 - \lambda^j \boldsymbol{g}_1^{jT} \boldsymbol{x}_1^j = 0, \quad (19)$$

$$2\|\tilde{\boldsymbol{x}}_2^j - \boldsymbol{x}_2^j\|^2 - \lambda^j \boldsymbol{g}_2^{jT} \boldsymbol{x}_2^j = 0, \quad (20)$$

from which the reprojection error for point j is given by

$$\|\tilde{\boldsymbol{x}}_1^j - \boldsymbol{x}_1^j\|^2 + \|\tilde{\boldsymbol{x}}_2^j - \boldsymbol{x}_2^j\|^2 = \frac{\lambda^j}{2} (\boldsymbol{g}_1^{jT} \boldsymbol{x}_1^j + \boldsymbol{g}_2^{jT} \boldsymbol{x}_2^j). \quad (21)$$

After replacing (18) in the previous equation, we obtain the following expression for the total reprojection error

$$\tilde{E}_n(\{F_i\}_{i=1}^n, \{(\tilde{\boldsymbol{x}}_1^j, \tilde{\boldsymbol{x}}_2^j)\}_{j=1}^N) \doteq \sum_{j=1}^{N} \frac{(\boldsymbol{g}_1^{jT} \boldsymbol{x}_1^j + \boldsymbol{g}_2^{jT} \boldsymbol{x}_2^j)^2}{\|[e_3]_\times \boldsymbol{g}_1^j\|^2 + \|[e_3]_\times \boldsymbol{g}_2^j\|^2} =$$

$$\sum_{j=1}^{N} \frac{(\boldsymbol{x}_1^{jT} (D\nu_n(\tilde{\boldsymbol{x}}_1^j))^T \mathcal{F}^T \nu_n(\tilde{\boldsymbol{x}}_2^j) + \boldsymbol{x}_2^{jT} (D\nu_n(\tilde{\boldsymbol{x}}_2^j))^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j))^2}{\|[e_3]_\times (D\nu_n(\tilde{\boldsymbol{x}}_1^j))^T \mathcal{F}^T \nu_n(\tilde{\boldsymbol{x}}_2^j)\|^2 + \|[e_3]_\times (D\nu_n(\tilde{\boldsymbol{x}}_2^j))^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j)\|^2}.$$

Since $\nu_1(\boldsymbol{x}) = \boldsymbol{x}$ and $D\nu_1(\boldsymbol{x}) = I$, by letting $n = 1$ in the above expression we notice that $\tilde{E}_n$ is a natural generalization of the well-known optimal function for estimating a *single* fundamental matrix $F \in \mathbb{R}^{3\times 3}$, which is given by [8]

$$\tilde{E}_1(F, \{(\tilde{\boldsymbol{x}}_1^j, \tilde{\boldsymbol{x}}_2^j)\}_{j=1}^N) = \sum_{j=1}^{N} \frac{(\boldsymbol{x}_1^{jT} F^T \tilde{\boldsymbol{x}}_2^j + \boldsymbol{x}_2^{jT} F \tilde{\boldsymbol{x}}_1^j)^2}{\|[e_3]_\times F^T \tilde{\boldsymbol{x}}_2^j\|^2 + \|[e_3]_\times F \tilde{\boldsymbol{x}}_1^j\|^2}. \tag{22}$$

Remark 3 *Notice that the optimal error $\tilde{E}_n$ has a very intuitive interpretation. If point j belongs to group i, then $\tilde{\boldsymbol{x}}_2^{jT} F_i \tilde{\boldsymbol{x}}_1^j = 0$. This implies that*

$$\begin{aligned}
\boldsymbol{g}_1^{jT} &= \nu_n(\tilde{\boldsymbol{x}}_2^j)^T \mathcal{F} D\nu_n(\tilde{\boldsymbol{x}}_1^j) = \frac{\partial}{\partial \tilde{\boldsymbol{x}}_1^j}\Big(\nu_n(\tilde{\boldsymbol{x}}_2^j)^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j)\Big) \\
&= \frac{\partial}{\partial \tilde{\boldsymbol{x}}_1^j}\left(\prod_{i=1}^{n} \tilde{\boldsymbol{x}}_2^{jT} F_i \tilde{\boldsymbol{x}}_1^j\right) = \sum_{i=1}^{n}\left(\prod_{\ell\neq i}^{n} \tilde{\boldsymbol{x}}_2^{jT} F_\ell \tilde{\boldsymbol{x}}_1^j\right)(\tilde{\boldsymbol{x}}_2^{jT} F_i) \\
&= \Big(\prod_{\ell\neq i}^{n} \tilde{\boldsymbol{x}}_2^{jT} F_\ell \tilde{\boldsymbol{x}}_1^j\Big)(\tilde{\boldsymbol{x}}_2^{jT} F_i), \\
\boldsymbol{g}_2^{jT} &= \Big(\prod_{\ell\neq i}^{n} \tilde{\boldsymbol{x}}_2^{jT} F_\ell \tilde{\boldsymbol{x}}_1^j\Big)(\tilde{\boldsymbol{x}}_1^{jT} F_i^T).
\end{aligned}$$

Therefore, the factor $\prod_{\ell\neq i}^{n}(\tilde{\boldsymbol{x}}_2^{jT} F_\ell \tilde{\boldsymbol{x}}_1^j)$ is in both the numerator and the denominator of $\tilde{E}_n$. Hence the contribution of point j to the error $\tilde{E}_n$ reduces to

$$\frac{(\tilde{\boldsymbol{x}}_2^{jT} F_i \boldsymbol{x}_1^j + \boldsymbol{x}_2^{jT} F_i \tilde{\boldsymbol{x}}_1^j)^2}{\|[e_3]_\times F_i^T \tilde{\boldsymbol{x}}_2^j\|^2 + \|[e_3]_\times F_i \tilde{\boldsymbol{x}}_1^j\|^2}, \tag{23}$$

which is the same as the contribution of point j to the optimal function for a single fundamental matrix F_i in (22). Therefore, the objective function $\tilde{E}_n$ is just a clever algebraic way of simultaneously writing a mixture of optimal objective functions for individual fundamental matrices into a single *objective function for* all *the fundamental matrices.*

We now derive an objective function that depends on the motion parameters only. As in the case of a single fundamental matrix [8], this can be done by considering the first order statistics of $\nu_n(\boldsymbol{x}_2^j)^T \mathcal{F} \nu_n(\boldsymbol{x}_1^j)$. It turns out that this is equivalent to setting $\tilde{\boldsymbol{x}}^j = \boldsymbol{x}^j$ in the above expression for $\tilde{E}_n$. Since $D\nu_n(\boldsymbol{x})\boldsymbol{x} = n\nu_n(\boldsymbol{x})$ we obtain the following function of the fundamental matrices $E_n(F_1, \ldots, F_n) \doteq$

$$\boxed{\sum_{j=1}^{N} \frac{4n^2(\nu_n(\boldsymbol{x}_2^j)^T \mathcal{F} \nu_n(\boldsymbol{x}_1^j))^2}{\|[e_3]_\times (D\nu_n(\tilde{\boldsymbol{x}}_1^j))^T \mathcal{F}^T \nu_n(\tilde{\boldsymbol{x}}_2^j)\|^2 + \|[e_3]_\times (D\nu_n(\tilde{\boldsymbol{x}}_2^j))^T \mathcal{F} \nu_n(\tilde{\boldsymbol{x}}_1^j)\|^2}}.$$

Notice that E_n is just a normalized version of the algebraic error (8). Furthermore, when $n = 1$, E_n reduces to the well-known objective function for estimating a single fundamental matrix F [8]

$$E_1(F) = \sum_{j=1}^{N} \frac{4(\boldsymbol{x}_2^{jT} F \boldsymbol{x}_1^j)^2}{\|[e_3]_\times F^T \boldsymbol{x}_2^j\|^2 + \|[e_3]_\times F \boldsymbol{x}_1^j\|^2}. \tag{24}$$

Notice that (24) can also be obtained by setting $\tilde{\boldsymbol{x}} = \boldsymbol{x}$ in (22). Therefore, the objective function $E_n(F_1, \ldots, F_n)$ is a natural generalization of well-known objective function $E_1(F)$ in single body structure from motion.

In summary, we have derived an objective function from which one can simultaneously estimate all the fundamental matrices $\{F_i\}_{i=1}^n$ using all the image pairs $\{(\boldsymbol{x}_1^j, \boldsymbol{x}_2^j)\}_{j=1}^N$, without prior segmentation of the image measurements. The fundamental matrices can be obtained by minimizing E_n using standard nonlinear optimization techniques. One can use the multibody linear algorithm in Section 3 to initialize the number of motions and the fundamental matrices.

Remark 4 (Pure translation and calibrated cases) *The case of linearly moving objects (Remark 1) or calibrated cameras can be easily handled by properly parameterizing the fundamental matrices and then minimizing over fewer parameters.*

5. Experimental Results

In this section, we evaluate the performance of the proposed algorithm with respect to the number of motions n and the amount of noise in the image measurements. We also test our approach by segmenting a real image sequence.

We first test the algorithm on synthetic data. We randomly pick $n = 1, 2, 3, 4$ collections of $N = 50n$ feature points and apply a different (randomly chosen) rigid body motion (R_i, T_i), with $R_i \in SO(3)$ the rotation and $T_i \in \mathbb{R}^3$ the translation. Zero-mean Gaussian noise with standard deviation (std) from 0 to 2.5 pixels is added to the first two entries of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, assuming an image size of 500×500 pixels. We run 1000 trials for each noise level. For each trial the error between the true motions $\{(R_i, T_i)\}_{i=1}^n$ and the estimates $\{(\hat{R}_i, \hat{T}_i)\}_{i=1}^n$ is computed as

$$\text{Rot. error} = \frac{1}{n}\sum_{i=1}^{n} \text{acos}\Big(\frac{\text{trace}(R_i \hat{R}_i^T) - 1}{2}\Big) \quad \text{(degrees)}.$$

$$\text{Trans. error} = \frac{1}{n}\sum_{i=1}^{n} \text{acos}\Big(\frac{T_i^T \hat{T}_i}{\|T_i\|\|\hat{T}_i\|}\Big) \quad \text{(degrees)}.$$

Figure 1 plots the mean error in rotation and translation as a function of noise. In all trials the number of motions was correctly estimated from equation (6) as $n = 1, 2, 3, 4$[1]. The algorithm gives an error of less than 3° for rotation and less than 10° for translation. As expected, the performance deteriorates as the number of motions n increases, especially for the rotation estimates.

We also tested the proposed approach by segmenting a real image sequence with $n = 3$ moving objects: a truck, a car and a box. Figure 2(a) shows the first frame of the sequence with the tracked features superimposed. We tracked a total of $N = 173$ point features: 44 "$\circ$" for the truck, 48 "$\square$" for the car and 81 "$\triangle$" for the box. We estimated the number of motions from (6) as $n = 3$ and minimized

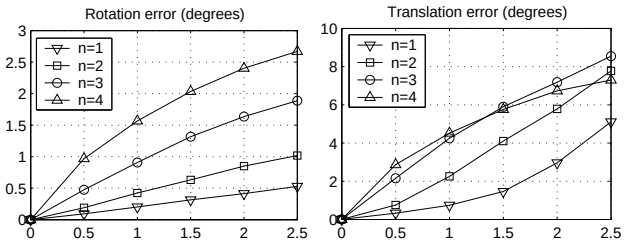
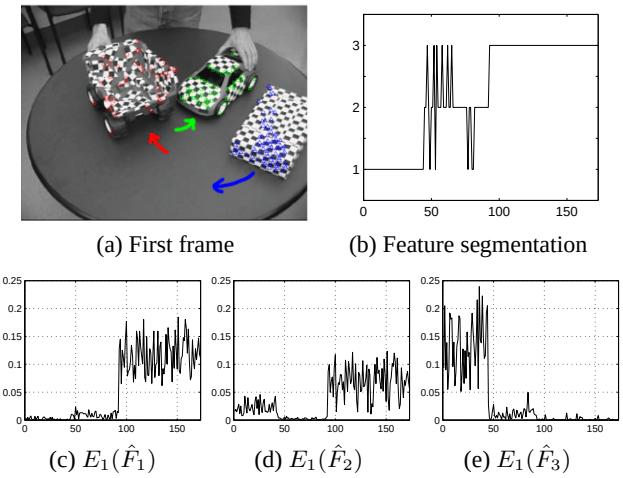


Figure 1: Error in the estimation of the rotation and translation as a function of noise in the image points (std in pixels).



(a) First frame (b) Feature segmentation

(c) $E_1(\hat{F}_1)$ (d) $E_1(\hat{F}_2)$ (e) $E_1(\hat{F}_3)$

Figure 2: 3-D segmentation of three independent motions.

$E_3(F_1, F_2, F_3)$ to obtain $(\hat{R}_i, \hat{T}_i)$. For comparison purposes, we estimated the ground truth motion (R_i, T_i) of each object by manually segmenting the feature points and then minimizing the standard error for single body structure from motion $E_1(F_i)$ in (24). The error in rotation was $1.5°$, $1.9°$ and $0.1°$ and the error in translation was $1.7°$, $1.8°$ and $0.4°$ for the truck, car and box, respectively.

In order to obtain the segmentation of the feature pairs, we computed the three reprojection errors $E_1(\hat{F}_i)$ for each feature pair as shown in Figures 2(c)-(e). Each feature pair was assigned to the motion $i = 1, 2, 3$ with the minimum error. Figure 2(b) plots the segmentation of the image points. There are no mismatches for motions 1 and 3. However 5 features corresponding to motion 2 are assigned to motion 1 and 6 features corresponding to motion 2 are assigned to motion 3. This is because the motion of the car was smaller and hence its reprojection error is small for all $\hat{F}_i$'s.

6. Conclusions

We presented a novel algorithm for optimally segmenting dynamic scenes containing multiple rigidly moving objects. Instead of iterating between feature segmentation and single body motion estimation, our approach eliminates the segmentation and directly optimizes over the motion parameters. We tested our approach on synthetic and real images.

References

[1] J. Costeira and T. Kanade. Multi-body factorization methods for motion analysis. In *IEEE International Conference on Computer Vision*, pages 1071–1076, 1995.

[2] X. Feng and P. Perona. Scene segmentation from 3D motion. In *International Conference on Computer Vision and Pattern Recognition*, pages 225–231, 1998.

[3] A. Fitzgibbon and A. Zisserman. Multibody structure and motion: 3D reconstruction of independently moving objects. In *European Conference on Computer Vision*, pages 891–906, 2000.

[4] M. Han and T. Kanade. Reconstruction of a scene with multiple linearly moving objects. In *International Conference on Computer Vision and Pattern Recognition*, volume 2, pages 542–549, 2000.

[5] M. Han and T. Kanade. Multiple motion scene reconstruction from uncalibrated views. In *IEEE International Conference on Computer Vision*, volume 1, pages 163–170, 2001.

[6] K. Kanatani. Motion segmentation by subspace separation and model selection. In *IEEE International Conference on Computer Vision*, volume 2, pages 586–591, 2001.

[7] K. Kanatani and C. Matsunaga. Estimating the number of independent motions for multibody motion segmentation. In *Asian Conference on Computer Vision*, pages 7–12, 2002.

[8] Y. Ma, J. Košecká, and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *International Journal of Computer Vision*, 44(3):219–249, 2001.

[9] A. Shashua and A. Levin. Multi-frame infinitesimal motion model for the reconstruction of (dynamic) scenes with multiple linearly moving objects. In *IEEE International Conference on Computer Vision*, volume 2, pages 592–599, 2001.

[10] S. Soatto and R. Brockett. Optimal and suboptimal structure from motion: local ambiguities and global estimates. In *International Conference on Computer Vision and Pattern Recognition*, pages 282–288, 1998.

[11] P. Sturm. Structure and motion for dynamic scenes - the case of points moving in planes. In *European Conference on Computer Vision*, pages 867–882, 2002.

[12] R. Szeliski and S. B. Kang. Recovering 3D shape and motion from image streams using non-linear least squares. *Journal of Visual Communication and Image Representation*, 5(1):10–28, 1994.

[13] P. H. S. Torr. Geometric motion segmentation and model selection. *Phil. Trans. Royal Society of London A*, 356(1740):1321–1340, 1998.

[14] R. Vidal, Y. Ma, S. Hsu, and S. Sastry. Optimal motion estimation from the multiview normalized epipolar constraint. In *IEEE International Conference on Computer Vision*, volume 1, pages 34–41, Vancouver, Canada, 2001.

[15] R. Vidal, Y. Ma, and S. Sastry. Generalized principal component analysis (GPCA). In *International Conference on Computer Vision and Pattern Recognition*, 2003.

[16] R. Vidal, Y. Ma, S. Soatto, and S. Sastry. Two-view multibody structure from motion. *International Journal of Computer Vision*, 2002. Submitted.

[17] R. Vidal and S. Sastry. Segmentation of dynamic scenes from image intensities. In *IEEE Workshop on Motion and Video Computing*, pages 44–49, 2002.

[18] J. Weng, N. Ahuja, and T. Huang. Optimal motion and structure estimation. *IEEE Transactions PAMI*, 9(2):137–154, 1993.

[19] L. Wolf and A. Shashua. Affine 3-d reconstruction from two projective images of independently translating planes. In *IEEE International Conference on Computer Vision*, pages 238–244, 2001.

[20] L. Wolf and A. Shashua. Two-body segmentation from two perspective views. In *International Conference on Computer Vision and Pattern Recognition*, pages 263–270, 2001.