Robust Unambiguous Parametrization of the Essential Manifold

Raghav Subbarao[†‡]
[†]ECE Department
Rutgers University
Piscataway, NJ 08854

Yakup Genc[‡] Peter Meer[†]
[‡]Real-time Vision and Modelling Department
Siemens Corporate Research
Princeton, NJ 08540

Abstract

Analytic manifolds were recently used for motion averaging, segmentation and robust estimation. Here we consider the epipolar constraint for calibrated cameras, which is the most general motion model for calibrated cameras and is encoded by the essential matrix. The set of all essential matrices forms the essential manifold. We provide a theoretical characterization of the geometry of the essential manifold and develop a parametrization which associates each essential matrix with a unique point on the manifold. Our work provides a more complete theoretical analysis of the essential manifold than previous work in this direction. We show the results of using this parametrization with real data sets, while previous work concentrated on theoretical analysis with synthetic data.

1. Introduction

Analytic manifolds have been used for various applications such as motion averaging [1, 7], motion segmentation [21, 22] and robust estimation [20]. The idea behind these algorithms is to consider a particular model, such as affine or rigid body motions, and the motion parameters correspond to single points on a manifold. Averaging or clustering these points over this manifold leads to motion averaging, motion segmentation *etc.*.

These techniques have been applied to motion models for which the manifold of parameters has been well studied. However, computer vision problems have geometric constraints such as the *epipolar constraint* [12], which are not common in other fields. The epipolar constraint encodes a relation between correspondences across two images of the same scene. In a calibrated setting, the epipolar constraint is parameterized by the *essential matrix*, a 3×3 matrix with certain algebraic properties [8, Sec.8.6.1]. The essential matrix represents the relative motion between two cameras [19], but due to the loss of depth information only the direction of translation can be recovered. It is possible to recover the relative camera geometry from the essential matrix, but there exists a four-fold ambiguity in this process and four different relative camera geometries correspond to each essential matrix [8, Sec.8.6.3].

Previous parametrizations of the essential manifold [11, 19] associate essential matrices with the rigid motions they encode. This was used for visual motion control [18, 19] and geometric optimization [14]. However, due to the four-fold ambiguity mentioned previously, each essential matrix corresponds to four different motions. The only way to choose among these motions is to enforce the *positive depth constraint* [18, 19], also known as *chierality* [8, Ch.20]. In the presence of mismatches and outliers, the use of image correspondence information can be a problem. For example, in a robust estimation problem where we are given a set of point correspondences between two images, some of the correspondences are mismatches. However, we have no knowledge of the true matches and mismatches. Using an incorrect match to enforce the positive depth constraint can lead to erroneous camera geometries.

We consider the problem of a *robust, unique* parametrization of the essential manifold. Rather than draw a parallel with rigid body motion, we use the algebraic properties of the essential matrix to parameterize the manifold. Each essential matrix will be associated with a *unique point* on the manifold, and there is no need to use any (possibly outlier) correspondence information to ensure a consistent local parametrization. This allows us to interpolate and average essential matrices directly without considering the relative camera geometries they encode.

In [9] a geometric optimization method over essential matrices was proposed using a similar idea. In [6], harmonic analysis was used for function optimization over the essential manifold using the same parametrization. However, the theoretical properties of the parametrization are not fully explored. Firstly, the essential manifold is not only a manifold but a *Riemannian manifold*. As we discuss later, this means that we can define a formal notion of distance between points. The previous work does not address or take advantage of this property. Secondly, we show that along with the Riemannian structure, the essential manifold belongs to a class of manifolds known as *ho-*

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

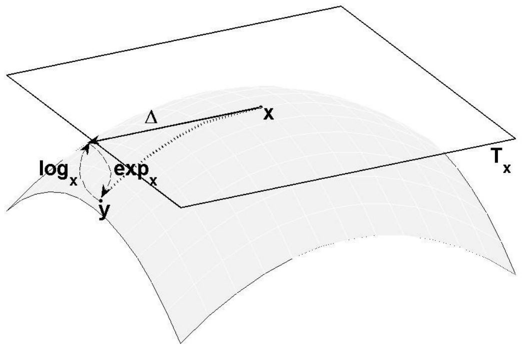


Figure 1. An illustration of a two-dimensional manifold and the tangent space, $\mathbf{T_X}$, at the point $\mathbf{X}$.

mogeneous spaces. This property makes the analysis considerably easier. For example, [5] proposes a methods to compute certain differential operators over homogeneous spaces. Finally, we exhibit results on real data. Both [6] and [9] treat the parametrization only as a theoretical tool. Although our main contributions are theoretical in nature, we show results on real data where we use the Riemannian structure of the essential manifold.

In Section 2 we briefly introduce the idea of analytic manifolds and the special orthogonal group. The essential manifold and its parametrization are discussed in Section 3. Section 4 discusses the nonlinear mean shift algorithm and how to use it for robust parameter estimation. The results are presented in Section 5.

2. Analytic Manifolds

A manifold, $\mathcal{M}$, is a surface which *locally* appears like Euclidean space. Formally, each point on $\mathcal{M}$ is associated with an open neighbourhood and a homeomorphism from the neighbourhood to an open set in Euclidean space. There is a smooth transition between the homeomorphisms in regions where the neighbourhoods overlap. The open sets in the above definition require the existence of a topology on $\mathcal{M}$. Analytic manifolds satisfy further conditions of smoothness [2, Sec.III.1]. We restrict ourselves to analytic manifolds and assume these conditions are satisfied.

The *tangent space* $\mathbf{T_X}$ at $\mathbf{X}$ can be thought of as the set of allowed instantaneous velocities for a point constrained to move on the manifold. The tangent space of a d-dimensional manifold is a d-dimensional vector space [2, Sec.IV.1]. We can define an inner product $g_\mathbf{X}$ on $\mathbf{T_X}$. This induces a norm for tangents $\Delta \in \mathbf{T_X}$ as $\|\Delta\|^2_\mathbf{X} = g_\mathbf{X}(\Delta, \Delta)$. The subscripts indicate the dependence of the inner product and norm on the point.

For a curve connecting two points, the derivative at each point lies in the tangent space at that point. The length of the curve is obtained by integrating the norm of the tangents along the curve [2, Sec.V.3]. The shortest curve joining two points on $\mathcal{M}$ is known as the *geodesic* and the length of the geodesic is the *intrinsic distance* between the points.

Tangents and geodesics are closely related. For each $\Delta \in \mathbf{T_X}$ there exists a unique geodesic starting at $\mathbf{X}$ with initial velocity Δ. This is captured by the *exponential map*, $exp_\mathbf{X}$, which maps Δ to the point on the manifold reached by this geodesic. In other words, the geodesic starting at $\mathbf{X}$ and ending at $exp_\mathbf{X}(\Delta)$ has initial velocity Δ. The inverse of the exponential map is the *logarithm map*, $log_\mathbf{X} = exp^{-1}_\mathbf{X}$ [2, Sec. VII.6]. These ideas are illustrated in Figure 1. The exponential maps tangents at $\mathbf{X}$ to points on the manifold and the logarithm maps points on the manifold to tangents at $\mathbf{X}$. Both exponential and logarithm operators vary as the point $\mathbf{X}$ moves. The specific forms of these operators depend on the manifold. The exponential is usually onto but not one-to-one. If many tangents satisfy $exp_\mathbf{X}(\Delta) = \mathbf{Y}$, $log_\mathbf{X}(\mathbf{Y})$ is the tangent with the smallest norm.

The *gradient* of a real function f defined on the manifold, is the *unique* tangent vector, $\nabla f \in \mathbf{T_X}$, satisfying

$$g_\mathbf{X}(\nabla f, \Delta) = \partial_\Delta f \tag{1}$$

for any $\Delta \in \mathbf{T_X}$, where ∂_Δ is the directional derivative along Δ. This is also known as the *differential* [2, Sec.V.1].

2.1. Special Orthogonal Group

A frequently occurring manifold is the set of 3D rotations, also known as the *special orthogonal group*, $\mathbf{SO}(3)$. This manifold consists of 3×3 orthogonal matrices with determinant one.

$$\mathbf{SO}(3) = \{\mathbf{X} \in \mathbb{R}^{3\times3} | \mathbf{X}\mathbf{X}^T = \mathbf{I}, det(\mathbf{X}) = 1\} \tag{2}$$

where, $\mathbf{I}$ is the 3×3 identity matrix. In fact, $\mathbf{SO}(3)$ belongs to the set of manifolds known as *Lie groups* [17] which have more algebraic structure than general manifolds. The group operation of $\mathbf{SO}(3)$ is matrix multiplication.

The Lie algebra $\mathfrak{so}(3)$ is the tangent space at the identity and consists of 3×3 skew-symmetric matrices of the form

$$[\omega]_\times = \begin{bmatrix} 0 & -\omega_z & \omega_y \\ \omega_z & 0 & -\omega_x \\ -\omega_y & \omega_x & 0 \end{bmatrix} \tag{3}$$

where the vector $\omega = [\omega_x \; \omega_y \; \omega_z]$ is the axis of rotation and $\|\omega\|$ is the magnitude of the rotation. The exponential operator is the matrix exponential and the logarithm operator is the matrix logarithm [17, 21, 22]. The structure of $\mathbf{SO}(3)$ allows us to compute the exponential using the *Rodriguez* formula [10, p.204]

$$exp([\omega]_\times) = \mathbf{I} + \frac{sin\|\omega\|}{\|\omega\|}[\omega]_\times + \frac{1 - cos\|\omega\|}{\|\omega\|^2}[\omega]^2_\times \, . \tag{4}$$

The logarithm is computed by inverting this relation. The intrinsic distance between $\mathbf{X}, \mathbf{Y} \in \mathbf{SO}(3)$ is

$$d(\mathbf{X}, \mathbf{Y}) = \|log(\mathbf{X}^{-1}\mathbf{Y})\|_F = \|log(\mathbf{X}^T\mathbf{Y})\|_F \tag{5}$$

since, for orthogonal matrices, $\mathbf{X}^{-1} = \mathbf{X}^T$ and $\| \cdot \|_F$ is the Frobenius norm.

Let $\mathbf{S}_z$ be the subgroup of rotations in $\mathbf{SO}(3)$ which leave the direction of the z-axis unchanged. It contains matrices of the form

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & det(\mathbf{A}) \end{bmatrix} \tag{6}$$

where, $\mathbf{A}$ is a 2×2 orthogonal matrix and $det(\mathbf{A}) = \pm 1$. The third diagonal element should be $det(\mathbf{A})$ to ensure $\mathbf{S}_z \subset \mathbf{SO}(3)$. Under the topology of $\mathbf{SO}(3)$, $\mathbf{S}_z$ is a *closed* subgroup of $\mathbf{SO}(3)$ [17, Sec.2.7].

3. The Essential Manifold

An essential matrix encodes the epipolar geometry for a set of calibrated cameras. Let p and q be the normalized coordinates of corresponding points and $\mathbf{Q}$ be the essential matrix. The essential constraint is

$$p^T \mathbf{Q} q = 0. \tag{7}$$

Let $\mathcal{E}$ denote the *essential space*, the set of all essential matrices. The essential space is an algebraic variety [15, 19] and a manifold of dimension six. Essential matrices have some further algebraic properties. If $\mathbf{Q} = \mathbf{U\Sigma V}^T$ is the singular value decomposition of a 3×3 matrix $\mathbf{Q}$, then [8, Sec.8.6.1]

$$\mathbf{Q} \in \mathcal{E} \Leftrightarrow \mathbf{\Sigma} = diag\{\lambda, \lambda, 0\}, \lambda \in \mathbb{R}^+ \tag{8}$$

i.e., an essential matrix has two equal, positive singular values and a zero singular value. The essential matrix is a homogeneous quantity and scaling does not change the geometry. We scale $\mathbf{Q}$ to ensure $\lambda = 1$ and define the *normalized essential space*, $\mathcal{E}_1$ as the set of 3×3 matrices with two unit singular values and one zero singular value

$$\mathbf{Q} \in \mathcal{E}_1 \Leftrightarrow \mathbf{\Sigma} = \mathbf{\Sigma}_1 \tag{9}$$

where, $\mathbf{\Sigma}_1 = diag\{1, 1, 0\}$.

Since the epipolar geometry depends on the relative pose of the cameras, it can be recovered from the essential matrix except for two ambiguities. Firstly, there is no scale information and the baseline between the cameras can only be recovered upto a scale. Secondly, four different relative camera geometries give rise to the same essential matrix [8, p.241] as shown in Figure 2. Usually, further image information is required to disambiguate the four geometries and choose the true geometry based on the *positive depth constraint*.

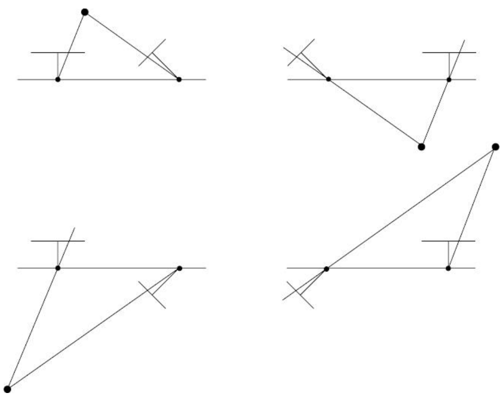


Figure 2. Four different camera geometries which give the same essential matrix. In each row the geometries differ by the sign of the direction of translation. Each column is a twisted pair. The image was taken from [8, p.241].

A common parametrization of the essential manifold is based on the fact that each relative camera geometry corresponds to a tangent of $\mathbf{SO}(3)$ with unit norm. The set of all tangents of a manifold forms a manifold known as the *tangent bundle*. Therefore, the essential manifold can be identified with the unit tangent bundle of $\mathbf{SO}(3)$ [14, 19]. Since each essential matrix corresponds to four different camera geometries, and each camera geometry corresponds to a different tangent of $\mathbf{SO}(3)$, this parametrization gives a four-fold covering of the essential manifold.

When comparing essential matrices by mapping them to tangents of $\mathbf{SO}(3)$ it is necessary to choose *consistently* from among the four possibilities. Otherwise, it is possible that dissimilar essential matrices are mapped to nearby points on the manifold. The only way to ensure a consistent parametrization is to use image point correspondences and enforce the positive depth constraint [18, 19]. However, while performing robust estimation we do not know if point matches are correct or not. Using a mismatch to disambiguate the motions could lead to erroneous results. We will use the algebraic properties of the normalized essential space to develop a unique local parametrization which is *not* dependent on image correspondence information.

3.1. Parametrizing the Essential Manifold

An alternate parametrization was proposed in [6]. This was based on a singular value decomposition of the essential matrix. We further develop this idea here and show that this parametrization gives a one-to-one correspondence between the points on the manifold and essential matrices. Furthermore, this parametrization makes the essential manifold a homogeneous space under the action of the group $\mathbf{SO}(3) \times \mathbf{SO}(3)$ and later we will use this to obtain geometrically meaningful Riemannian metrics for the essential manifold.

Consider $\mathbf{Q} \in \mathcal{E}_1$ with singular value decomposition $\mathbf{U}\mathbf{\Sigma}_1\mathbf{V}^T$, where $\mathbf{U}$ and $\mathbf{V}$ are orthogonal and $det(\mathbf{U}), det(\mathbf{V}) = \pm 1$. As the third singular value is zero, we can change the sign of the third columns of $\mathbf{U}$ and $\mathbf{V}$ to ensure $det(\mathbf{U}), det(\mathbf{V}) = 1$ without changing the SVD.

Since $\mathbf{SO}(3)$ is a Lie group, the manifold $\mathbf{SO}(3) \times \mathbf{SO}(3)$ is also a Lie group with the topology and group operation inherited from $\mathbf{SO}(3)$ [17, Sec.4.3]. We define the mapping

$$\Phi : \mathbf{SO}(3) \times \mathbf{SO}(3) \rightarrow \mathcal{E}_1 \tag{10}$$

which maps $(\mathbf{U}, \mathbf{V}) \in \mathbf{SO}(3) \times \mathbf{SO}(3)$ to $\mathbf{U}\mathbf{\Sigma}_1\mathbf{V}^T \in \mathcal{E}_1$. The inverse mapping from $\mathcal{E}_1$ to $\mathbf{SO}(3) \times \mathbf{SO}(3)$ is not well defined as there is one degree of freedom (dof) in choosing the basis of the space spanned by the first two columns of $\mathbf{U}$ and $\mathbf{V}$. A rotation of the first two columns of $\mathbf{U}$ can be offset by a rotation of the first two columns of $\mathbf{V}$, such that $\mathbf{U}\mathbf{\Sigma}_1\mathbf{V}^T$ does not change. Consider $\mathbf{X}, \mathbf{Y} \in \mathbf{S}_z$ such that

$$\mathbf{X} = \begin{bmatrix} \mathbf{A} & \mathbf{0} \\ \mathbf{0} & det(\mathbf{A}) \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \pm\mathbf{A} & \mathbf{0} \\ \mathbf{0} & det(\mathbf{A}) \end{bmatrix}.$$

and $\mathbf{A}\mathbf{A}^T = \pm\mathbf{I}$. Then, substitution gives

$$\mathbf{U}\mathbf{X}\mathbf{\Sigma}_1\mathbf{Y}^T\mathbf{V}^T = \mathbf{U}\begin{bmatrix} \pm\mathbf{A}\mathbf{A}^T & \mathbf{0} \\ \mathbf{0} & 0 \end{bmatrix}\mathbf{V}^T = \pm\mathbf{U}\mathbf{\Sigma}_1\mathbf{V}^T \tag{11}$$

which leaves the essential matrix unchanged and Φ maps $(\mathbf{U}\mathbf{X}, \mathbf{V}\mathbf{Y}) \in \mathbf{SO}(3) \times \mathbf{SO}(3)$ to the same point in $\mathcal{E}_1$.

Let $\mathbf{H}_\Phi$ be the symmetry group which leaves Φ invariant. It consists of elements which leave the third columns of $\mathbf{U}$ and $\mathbf{V}$ unchanged, and rotate the first two columns by angles which differ by $k\pi, k \in \mathbb{Z}$.

$$\mathbf{H}_\Phi = \{(\mathbf{R}_1, \mathbf{R}_2) | \mathbf{R}_1, \mathbf{R}_2 \in \mathbf{S}_z, \mathbf{R}_1^T\mathbf{R}_2 = \mathbf{R}_z(k\pi)\} \tag{12}$$

where, $\mathbf{R}_z(k\pi)$ denotes a rotation by $k\pi$ around the z-axis. Since $\mathbf{S}_z$ is a closed subgroup of $\mathbf{SO}(3)$, it can be shown that $\mathbf{H}_\Phi$ is a closed subgroup of $\mathbf{SO}(3) \times \mathbf{SO}(3)$.

The manifold $\mathcal{E}_1$ is identified with the manifold $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$. This notation means that elements of $\mathbf{SO}(3) \times \mathbf{SO}(3)$ which differ by group multiplication by an element in $\mathbf{H}_\Phi$ are considered to be the same on $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$. Two elements differing only by multiplication by an element in $\mathbf{H}_\Phi$ are said to be in the same *equivalence class*, and (11) shows that such elements represent the same essential matrix. Multiplication of $(\mathbf{U}, \mathbf{V})$ by elements of $\mathbf{H}_\Phi$ generates the equivalence class of $(\mathbf{U}, \mathbf{V})$. Since $\mathbf{H}_\Phi$ is a closed subgroup of $\mathbf{SO}(3) \times \mathbf{SO}(3)$, we can use the *closed subgroup theorem* [17, Sec.2.7] to show that $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$ inherits topology and local coordinates from $\mathbf{SO}(3) \times \mathbf{SO}(3)$ [17, Sec.4.2]. Each element of $\mathcal{E}_1$ corresponds to a *unique* element in $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$ or a single equivalence class in $\mathbf{SO}(3) \times \mathbf{SO}(3)$.

3.2. Vertical and Horizontal Spaces

The manifold $\mathbf{SO}(3) \times \mathbf{SO}(3)$ consists of two copies of $\mathbf{SO}(3)$ and the tangent space of $\mathbf{SO}(3) \times \mathbf{SO}(3)$ will consist of two copies of the tangent space of $\mathbf{SO}(3)$. Since $\mathbf{SO}(3)$ has three-dimensional tangent spaces, $\mathbf{SO}(3) \times \mathbf{SO}(3)$ will have six-dimensional tangent spaces. Consider $(\mathbf{U}, \mathbf{V}) \in \mathbf{SO}(3) \times \mathbf{SO}(3)$ and a tangent represented as a six vector

$$\Delta = \begin{bmatrix}\mathbf{u}^T \ \mathbf{v}^T\end{bmatrix}^T \tag{13}$$

where, $\mathbf{u} = \begin{bmatrix}u_x \ u_y \ u_z\end{bmatrix}^T$ and $\mathbf{v} = \begin{bmatrix}v_x \ v_y \ v_z\end{bmatrix}^T$. The exponential for $\mathbf{SO}(3) \times \mathbf{SO}(3)$ is computed by performing the exponential of $\mathbf{SO}(3)$ twice, once each for $\mathbf{U}$ and $\mathbf{V}$

$$exp_{(\mathbf{U},\mathbf{V})}(\Delta) = (\ \mathbf{U}exp([\mathbf{u}]_\times), \mathbf{V}exp([\mathbf{v}]_\times)\ ) \tag{14}$$

where, the exp on the right represents the matrix exponential computed by the Rodriguez formula (4) and $[\cdot]_\times$ is defined by (3). The first three elements of the tangent vector correspond to $\mathbf{U}$ and the last three to $\mathbf{V}$. This ordering is equivalent to choosing a basis for the tangent space.

The tangent space of $\mathbf{SO}(3) \times \mathbf{SO}(3)$ can be divided into two complementary subspaces. The *horizontal space* contains tangents of the form

$$[u_x \ \ u_y \ \ u_z \ \ v_x \ \ v_y \ \ -u_z], \qquad \|u_z\| < \pi/2. \tag{15}$$

The *vertical space* consists of tangents of the form

$$[0 \ \ 0 \ \ u_z \ \ 0 \ \ 0 \ \ k\pi + u_z] \tag{16}$$

which lie in the Lie algebra of $\mathbf{H}_\Phi$ [4]. When $k = 0$, the vertical and horizontal spaces form complementary subspaces around the origin of the tangent space. Moving along geodesics defined by tangents in the vertical space is equivalent to multiplying by elements of $\mathbf{H}_\Phi$ and leaves the equivalence class unchanged. Tangents in the horizontal space are tangent to the equivalence class and all tangents of $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$ must lie in the horizontal space of $\mathbf{SO}(3) \times \mathbf{SO}(3)$. Given a tangent in the horizontal space, its exponential can be computed by (14) to get an element in another equivalence class, representing a different essential matrix.

Let $(\mathbf{U}, \mathbf{V})$ and $(\hat{\mathbf{U}}, \hat{\mathbf{V}})$ represent two elements of $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$. These can be any points in their respective equivalence classes. The logarithm operator for $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$ should give a tangent in the horizontal space. To do this we first compute the logarithm on the manifold $\mathbf{SO}(3) \times \mathbf{SO}(3)$. Define

$$\delta\mathbf{U} = \mathbf{U}^T\hat{\mathbf{U}} \quad \delta\mathbf{V} = \mathbf{V}^T\hat{\mathbf{V}}. \tag{17}$$

Taking the matrix logarithms of $\delta\mathbf{U}$ and $\delta\mathbf{V}$, and rearranging the elements into a six-vector, we get

$$[u_x \ u_y \ u_z \ v_x \ v_y \ v_z]^T \tag{18}$$

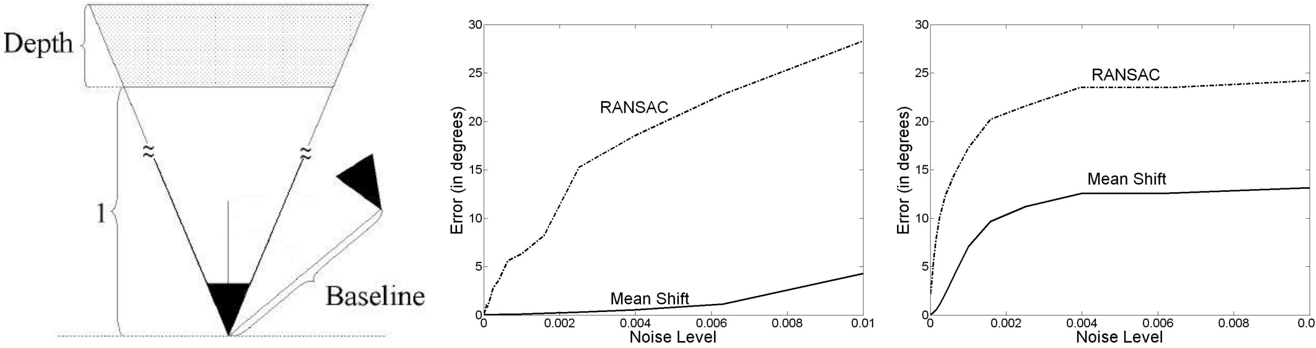


Figure 3. Synthetic data. A scene setup is shown on the left. The *middle image* compares the performance of RANSAC and mean shift when the baseline is perpendicular to the viewing direction. The *right image* compares performance when the baseline and viewing direction are parallel. The abscissa is the level of image noise in normalized coordinates, and a value of 0.01 is about 5 pixels in a 640×480 image.

which lies in the tangent space of $\mathbf{SO}(3) \times \mathbf{SO}(3)$. Since $(\mathbf{U}, \mathbf{V})$ and $(\hat{\mathbf{U}}, \hat{\mathbf{V}})$ are arbitrary elements of their equivalence classes, it is not necessary that this vector lie in the horizontal space. We need to remove the component lying in the vertical space. Using Givens rotations [8, App.3] $\delta\mathbf{U}$ and $\delta\mathbf{V}$ are decomposed into rotations around the z-axis and rotations around axes in the xy-plane. Now, $(\mathbf{U}, \mathbf{V})$ is moved using z-rotations differing by $k\pi$, according to (14), so that on recomputing $\delta\mathbf{U}$ and $\delta\mathbf{V}$, they have opposite z-rotations less than $\pi/2$. This can be done in a single step and ensures that for the new $\delta\mathbf{U}$ and $\delta\mathbf{V}$, $u_z \approx -v_z$ upto a few degrees. Due to the nonlinearity of the manifold $u_z = -v_z$ will not hold exactly. This can be improved by moving $(\mathbf{U}, \mathbf{V})$ along tangents of the form

$$[0\ \ 0\ \ (u_z + v_z)/2\ \ 0\ \ 0\ \ (u_z + v_z)/2]^T \qquad (19)$$

and recomputing $\delta\mathbf{U}$ and $\delta\mathbf{V}$. The tangents of (19) lie in the vertical space and do not change the equivalence class of $(\mathbf{U}, \mathbf{V})$. After the initial step with Givens rotations, $u_z + v_z$ is very small. Three or four iterations generally give an acceptable accuracy of the order of 10^{-4}. At convergence we obtain the *log*, which is a six-vector of the form

$$[u_x\ u_y\ u_z\ v_x\ v_y\ \ -u_z] \qquad (20)$$

pointing from one equivalence class to the other. The intrinsic distance between $(\mathbf{U}, \mathbf{V})$ and $(\hat{\mathbf{U}}, \hat{\mathbf{V}})$ is given by the norm of the five dimensional vector

$$d\left((\mathbf{U}, \mathbf{V}), (\hat{\mathbf{U}}, \hat{\mathbf{V}})\right) = \|[u_x\ u_y\ u_z\ v_x\ v_y]\|. \qquad (21)$$

We repeat that an essential matrix, irrespective of the camera geometry it represents, is a single point on the manifold $\mathbf{SO}(3) \times \mathbf{SO}(3)/\mathbf{H}_\Phi$.

4. Nonlinear Mean Shift

Consider a manifold with a metric d. Given n points on the manifold, $\mathbf{X}_i, i = 1, \ldots, n$, the *kernel density estimate* with *profile* k and *bandwidth* h is

$$\hat{f}_k(\mathbf{X}) = \frac{c_{k,h}}{n} \sum_{i=1}^{n} k\left(\frac{d^2(\mathbf{X}, \mathbf{X}_i)}{h^2}\right). \qquad (22)$$

The bandwidth h can be included in the distance as a parameter. However, written in this form, the bandwidth gives a parameter which can be used to tune performance. If the manifold is Euclidean with a Euclidean distance metric, (22) reduces to a Euclidean kernel density estimate [3].

Taking the gradient of $\hat{f}_k$ at $\mathbf{X}$,

$$\begin{aligned} \nabla \hat{f}_k(\mathbf{X}) &= \frac{1}{n} \sum_{i=1}^{n} \nabla k\left(\frac{d^2(\mathbf{X}, \mathbf{X}_i)}{h^2}\right) \qquad (23) \\ &= -\frac{1}{n} \sum_{i=1}^{n} g\left(\frac{d^2(\mathbf{X}, \mathbf{X}_i)}{h^2}\right) \frac{\nabla d^2(\mathbf{X}, \mathbf{X}_i)}{h^2} \end{aligned}$$

where, $g(z) = -k'(z)$. The gradient of $\nabla d^2(\mathbf{X}, \mathbf{X}_i)$ is taken with respect to $\mathbf{X}$. It was shown in [21] that for Lie groups, $\nabla d^2(\mathbf{X}, \mathbf{X}_i) = -log_{\mathbf{X}}(\mathbf{X}_i)$. The nonlinear mean shift vector is

$$\mathbf{m}_h(\mathbf{X}) = \frac{\sum_{i=1}^{n} log_{\mathbf{X}}(\mathbf{X}_i) g\left(\frac{d^2(\mathbf{X}, \mathbf{X}_i)}{h^2}\right)}{\sum_{i=1}^{n} g\left(\frac{d^2(\mathbf{X}, \mathbf{X}_i)}{h^2}\right)} \qquad (24)$$

The operations in (24) are well defined. The gradient terms, $\nabla d^2(\mathbf{X}, \mathbf{X}_i)$ lie in the tangent space $\mathbf{T}_\mathbf{X}$, and the kernel terms $g(d^2(\mathbf{X}, \mathbf{X}_i)/h^2)$ are scalars. The mean shift vector is a weighted average of tangent vectors, and lies in $\mathbf{T}_\mathbf{X}$. The iteration moves the point along the geodesic defined by the mean shift vector. The nonlinear mean shift iteration is

$$\mathbf{X}^{(j+1)} = exp_{\mathbf{X}^{(j)}}\left(\mathbf{m}_h(\mathbf{X}^{(j)})\right). \qquad (25)$$

The iteration ([25]), moves the current mode estimate $\mathbf{X}^{(j)}$ along the geodesic defined by the mean shift vector, to get the next updated estimate, $\mathbf{X}^{(j+1)}$. For the essential manifold the exp and log operators and the distance can be computed as discussed in Section 3.2.

4.1. Robust Estimation and Segmentation

The robust estimation algorithm based on nonlinear mean shift was proposed in [21, 22] and its properties are discussed in [20]. A similar procedure can also be used for motion segmentation.

The input consists of a set of point matches. The algorithm proceeds in two stages. In the first stage, the matches are randomly sampled to generate *elemental subsets*. An elemental subset consists of the minimum number of points required to specify a hypotheses. For essential matrices, an elemental subset consists of five point matches and the algorithm of [16] is used to generate the hypotheses. Each elemental subset generates multiple (up to 11) solutions for the essential matrix. These solutions correspond to *different* essential matrices and should not be confused with the fourfold ambiguity of the camera geometries corresponding to a single essential matrix. The hypothesis generation can be improved by a *validation* step which reduces computation in the second stage [22]. In the second stage, the parameter estimates are clustered using nonlinear mean shift. In the presence of multiple motions mean shift will find multiple modes and the number of dominant modes should be the number of motions [21]. The position of the mode is the essential matrix being estimated. Given the essential matrix, the inliers can be obtained as discussed in [21]. Briefly, the residual errors are computed for all the points and the first minima on either side of zero are found. Points with error lying in this window are declared inliers and the rest of the points are outliers.

5. Experimental Results

The behavior of the algorithm varies with the bandwidth. For each experiment, we chose the value based on the scene and the level of noise. All experiments were conducted on a Pentium D (2.79 GHz). Running mean shift for for 1000 points typically takes about 10s.

5.1. Synthetic Data

We tested our procedure on synthetic data under the same conditions as in [16]. In [16] the performance under noise of the 5-point hypothesis generation method was compared to other minimal case algorithms. The default synthetic data consisted of a scene with depth of 0.5 at a distance of 1 unit from the first camera. The baseline between the two cameras was taken to be 0.1 units. This data set reflects challenging, realistic conditions. Image noise of increasing levels was added to the correspondences and the performance of RANSAC was compared to mean shift. The estimated essential matrices are used to get the rotation and translation. Since the translation direction is much more sensitive than the rotation estimates, the error is the deviation of the estimated direction of translation from the true direction [16].

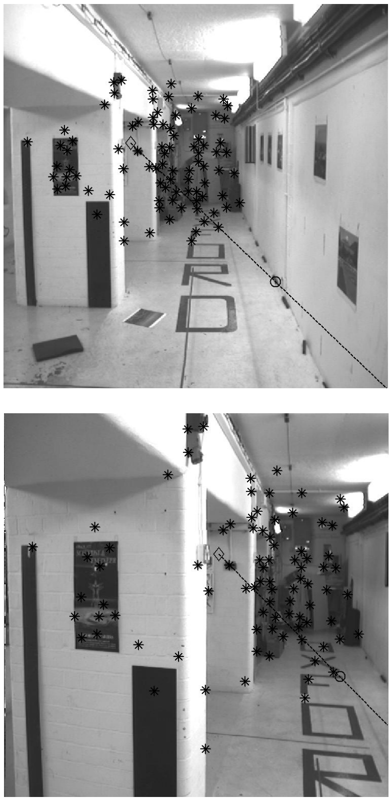


Figure 4. *Corridor* Images. The true essential matrix was the most dominant mode. See text for further details.

The basic scene setup for the synthetic experiments is shown in Figure 3. We show results under varying directions and different levels of noise. The noise was added to the normalized image coordinates. All normalized coordinates lie between -0.5 and 0.5 and a noise level of 0.01 corresponds to a standard deviation of about 5 pixels in a 640×480 image. The middle figure compares RANSAC and mean shift when the baseline is perpendicular to the viewing direction. As the noise levels increase averaging the hypothesis offers a clear advantage over choosing the best hypothesis. The right figure compares the results when the baseline is parallel to the viewing direction. In this case, both errors do not increase beyond a certain level due to the geometry, but mean shift clearly does better than RANSAC.

5.2. Robust Estimation: *Corridor* Images

We used the first and last images of the *Corridor* sequence from Oxford to test the robust estimation algorithm.

Using SIFT [13] we obtain 130 matches of which 87 were inliers. We sampled elemental subsets and generated 500 essential matrix hypotheses using [16]. Mean shift was run with a bandwidth of 0.1. The dominant mode has a support (kernel density at the point) of 0.36, which is two orders of magnitude above the next mode with a support of 0.005. All 87 inliers were correctly identified and one outlier was misclassified as an inlier. This happened because the mismatch satisfied the essential constraint. The results are shown in Figure 4. The two frames are shown with the inliers marked as stars. The misclassified outlier is shown as a circle and the epipoles as diamonds. In the first image, the epipolar line of the mismatch passes through the top right corner of the letter ”F” on the floor. This point is matched with the top right corner of the ”F” in the second frame and the epipolar constraint is satisfied. Comparing the returned essential matrix with the ground truth, we found the estimate to be accurate. Comparing RANSAC to ground truth we find that mean shift does much better.

5.3. Robust Estimation: *Parking lot* Images

Another example of robust estimation is shown in Figure 5. The camera was calibrated offline and points were matched across the images and the essential matrix was estimated. Of the 126 point matches, 64 were inliers. There was no ground truth available, so the essential matrix computed using only the inliers was taken as ground truth. The robust estimate returned by the mean shift with bandwidth 0.001, was very close to the true essential matrix. All the inliers were correctly identified and 7 outliers were misclassified as inliers. Again, this is because the mismatches were such that they satisfied the essential constraint. The returned point matches are shown in Figure 5. The true inliers are shown as stars, the misclassified outliers are drawn as circles and the epipoles as diamonds. Consider, the top most circle in the top image. It is matched to the left most point on the bottom image. We can see that the epipolar line in the bottom image passes through the top right corner of the building, which is where the correct match should be. Therefore, though the points are mismatched, the essential constraint is satisfied and the match is declared an inlier. Like before, mean shift does better than RANSAC.

5.4. Motion Segmentation: *Lab* Images

The two images used for motion segmentation are shown in Figure 6. The toy cars move together and have the same essential matrix, while the book has a different essential matrix. Using SIFT, and removing points in the background as having zero displacement, we get 100 point matches with 39 on the book and 42 on the cars. A 1000 hypotheses were generated and clustered with a bandwidth of 0.001. The clustering returns two dominant modes. The inliers for each were found like in the previous examples. If a point was declared an inlier for both motions, it was assigned to the motion which gave a lower absolute error. The points are shown in Figure 6.

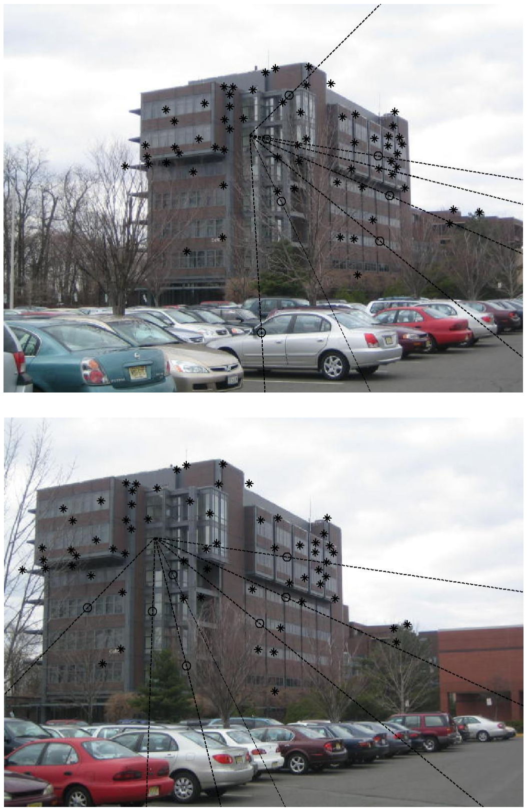
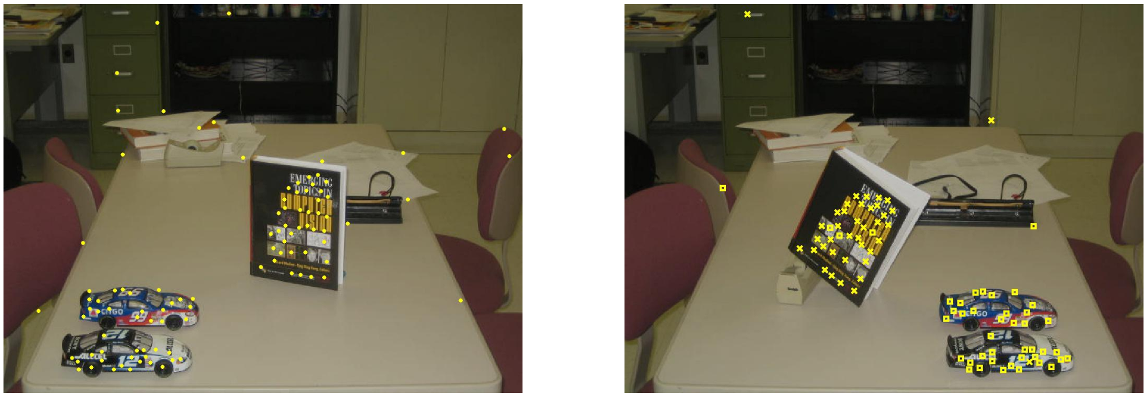


Figure 5. *Parking lot* Images. The scene contains 126 matches with 64 inliers. All inliers were correctly detected and 7 outliers were misclassified as inliers as they satisfy the essential constraint.

The results are tabulated below the figure. The first mode is due to the cars and the second mode is due to the book. The table on the left indicates the number of points which converge to each mode and the kernel density at the mode. The third mode has far fewer hypotheses converging to it, when compared to the first two modes. The table on the right shows the inlier-outlier classification. The first row indicates that of the 39 inliers on the cars, 36 have been correctly classified, one has been assigned to the book and two have been declared outliers. Similarly, the second row is about the book and the third row represents the outliers.

6. Conclusion

We propose a new parametrization of the essential manifold based on the algebraic properties of normalized essential matrices. We show that the essential manifold is a Riemannian manifold and also a homogeneous space. This allows us to define geometrically meaningful distances be-



	mot.hyp.	kde
$\mathbf{M}_1$	459	0.0215
$\mathbf{M}_2$	409	0.0051
M_3	92	0.0026

	$\mathbf{M}_1$	$\mathbf{M}_2$	*Out*
$\mathbf{M}_1$	36	1	2
$\mathbf{M}_2$	3	38	2
Out	0	3	15

Figure 6. *Motion Segmentation.* In the *left figure* all the points are plotted, and on the *right figure* only the returned inliers for the two motions are shown. The table on the left contains the properties of the first three modes. Only the first two modes correspond to motions. The table on the right compares the results with the ground truth.

tween essential matrices. Previous methods suffer from the disadvantage that each essential matrix corresponds to multiple points on the manifold. Choosing consistent local neighborhoods requires image correspondence information which is a problem in the presence of mismatches. For our method, each essential matrix corresponds to a unique point on the manifold and we can choose consistent local neighborhoods without using image correspondence information.

References

[1] E. Begelfor and M. Werman. How to put probabilities on homographies. *IEEE Trans. PAMI*, 27(10):1666–1670, 2005. 1

[2] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 2002. 2

[3] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Trans. PAMI*, 24:603–619, May 2002. 5

[4] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. 4

[5] R. Ferreira and J. Xavier. Hessian of the Riemannian squared-distance function on connected locally symmetric spaces with applications. In *Controlo 2006, 7th Portuguese Conference on Automatic Control*, 2006. 2

[6] C. Geyer, R. Bajcsy, and S. Sastry. Euclid meets Fourier: Applying harmonic analysis to essential matrix estimation in omnidirectional cameras. In *Proc.of Workshop on Omnidirectional Vision, Camera Networks and Non-classical Cameras*, 2004. 1, 2, 3

[7] V. M. Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* Washington, DC, volume I, pages 684–691, 2004. 1

[8] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000. 1, 3, 5

[9] U. Helmke, K. Huper, P. Lee, and J. Moore. Essential matrix estimation using Gauss-Newton iterations on a manifold. *International J. of Computer Vision*, 2007. 1, 2

[10] K. Kanatani. *Group Theoretical Methods in Image Understanding*. Springer-Verlag, 1990. 2

[11] J. Kosecka, Y. Ma, and S. Sastry. Optimization criteria, sensitivity and robustness. In B. Triggs, A. Zisserman, and R. Szelisky, editors, *Vision Algorithms: Theory and Practice*, pages 168–182. Springer, 2000. 1

[12] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981. 1

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International J. of Computer Vision*, 60(2):91–110, 2004. 7

[14] Y. Ma, J. Kosecka, and S. Sastry. Optimization criteria and geometric algorithms for motion and structure estimation. *International J. of Computer Vision*, 44(3):219–249, 2001. 1, 3

[15] S. J. Maybank. *Theory of Reconstruction from Image Motion*. Springer-Verlag, 1992. 3

[16] D. Nister. An efficient solution to the five-point relative pose problem. *IEEE Trans. PAMI*, 26(6):756–770, 2004. 6, 7

[17] W. Rossmann. *Lie Groups: An Introduction through Linear Groups*. Oxford University Press, 2003. 2, 3, 4

[18] S. Soatto, R. Frezza, and P. Perona. Recursive motion estimation on the essential manifold. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* San Francisco, CA, pages 428–433, 1993. 1, 3

[19] S. Soatto, P. Perona, R. Frezza, and G. Picci. Motion estimation via dynamic vision. In *Proc. European Conf. on Computer Vision,* Stockholm, Sweden, volume II, pages 61–72, 1994. 1, 3

[20] R. Subbarao, Y. Genc, and P. Meer. Nonlinear mean shift for robust pose estimation. In *8th IEEE Workshop on Applications of Computer Vision,* Austin, TX, February 2007. 1, 6

[21] R. Subbarao and P. Meer. Nonlinear mean shift for clustering over analytic manifolds. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition,* New York, NY, volume I, pages 1168–1175, 2006. 1, 2, 5, 6

[22] O. Tuzel, R. Subbarao, and P. Meer. Simultaneous multiple 3D motion estimation via mode finding on Lie groups. In *Proc. 10th Intl. Conf. on Computer Vision,* Beijing, China, volume 1, pages 18–25, 2005. 1, 2, 6