International Journal of Computer Vision 75(3), 311–327, 2007

DOI: 10.1007/s11263-007-0035-2

Correspondence-free Structure from Motion

AMEESH MAKADIA*,†
University of Pennsylvania, Philadelphia, PA 19104
makadia@cis.upenn.edu

CHRISTOPHER GEYER‡
Carnegie Mellon University, Pittsburgh, PA 15213
cgeyer@cs.cmu.edu

KOSTAS DANIILIDIS*
University of Pennsylvania, Philadelphia, PA 19104
kostas@cis.upenn.edu

Received January 4, 2006; Accepted January 2, 2007

First online version published in February, 2007

Abstract. We present a novel approach for the estimation of 3D-motion directly from two images using the Radon transform. The feasibility of any camera motion is computed by integrating over all feature pairs that satisfy the epipolar constraint. This integration is equivalent to taking the inner product of a similarity function on feature pairs with a Dirac function embedding the epipolar constraint. The maxima in this five dimensional motion space will correspond to compatible rigid motions. The main novelty is in the realization that the Radon transform is a filtering operator: If we assume that the similarity and Dirac functions are defined on spheres and the epipolar constraint is a group action of rotations on spheres, then the Radon transform is a correlation integral. We propose a new algorithm to compute this integral from the spherical Fourier transform of the similarity and Dirac functions. Generating the similarity function now becomes a preprocessing step which reduces the complexity of the Radon computation by a factor equal to the number of feature pairs processed. The strength of the algorithm is in avoiding a commitment to correspondences, thus being robust to erroneous feature detection, outliers, and multiple motions.

Keywords: motion estimation, structure from motion, registration, harmonic analysis, correspondence-free motion

1. Introduction

Estimation of 3D-motion from two calibrated views has been exhaustively studied in the case where optical flow or feature correspondences are given and the scene is rigid. Algorithms working over multiple frames yield high-quality motion trajectories and reconstructions when feature matches are cleaned through outlier rejection and motions independent of the camera are excluded. These outlier rejection and segmentation steps are subject to the fundamental coupling of data association and estimation: if we knew the motion estimate, data association would be trivial; if we knew the data association, motion estimation would be easier. Resistance to outliers and independent motions pose severe practical limitations to the wide application of structure from motion as a navigation tool, visual GPS, or a camera tracker.

In this paper, we propose a novel approach for structure from motion applicable in the presence of large motions and many irrelevant features resulting from reduced overlap of the fields of view. Our approach is based on the naive principle that an exhaustive search over all possible correspondence configurations for all motion hypotheses would yield all 3D-motions compatible with these two views. Such a search is intractable

*The authors are grateful for support through the following grants: NSF-IIS-0083209, NSF-IIS-0121293, NSF-EIA-0324977, NSF-CNS-0423891, NSF-IIS-0431070, and ARO/MURI DAAD19-02-1-0383.
†Correspondence author.
‡The author is grateful for the generous support of the ARO MURI program (DAAD-19-02-1-0383) while at U. C. Berkeley.

when we use a large field of view in an arbitrary, possibly unstructured environment with thousands of features.

The contribution of this paper is in the re-formulation of this Hough-reminiscent approach as a filtering problem: Assuming a similarity function between any two features in the first and second view, we convolve this function with a kernel that checks the compatibility of a correspondence pair with the epipolar constraint for a given motion hypothesis. The resulting integral is a Radon transform known from computer tomography where a material density is integrated over a ray path. In our case, this path is the subset of the cross product of all features that satisfies the epipolar constraint.

The question is: can we efficiently compute this integral avoiding the combinatorially infeasible summation over all correspondences compatible with the epipolar constraint? The answer is yes, because this is a convolution integral and we can compute it through multiplication in the Fourier domain. The final motion space is obtained through a five dimensional inverse rotational Fourier transform on the motion parameters. An exhaustive search finds the maxima corresponding to rigid motions. The number of spherical Fourier coefficients preserved determines the resolution of the motion space. Obviously, the approach can work on arbitrarily large motions.

We present a complete end-to-end system, from images to motion parameters where the only tuning parameter is the coupled resolution of the image and the motion space. We extract SIFT features (Lowe, 2004) for which we define their similarity function proportional to the Euclidean norm of the attribute vectors and we compute the spherical harmonics of the similarity function as the input to the correlation integral. In the experiments, we use as input hemispherical omnidirectional images. A projective plane can always be mapped to the sphere and the field of view has to be large for any structure from motion algorithm to succeed (Daniilidis and Spetsakis, 1996; Oliensis, 2000). The results on real sequences are compared to a robust estimation of the essential matrix using RANSAC. Before continuing with the related work we summarize the main contributions of this paper:

- We propose a new integral transform that maps a similarity function between two calibrated images to the strength of a motion hypothesis without assuming any correspondences.
- We show that this Radon/Hough transform can be written as a convolution/correlation integral which can be computed from the spherical harmonic coefficients of the image similarity function much faster than computing directly the Hough transform.

The inspiring idea of this work has first been drafted in Geyer et al. (2004) where a Hough transform is computed on the essential manifold. A short version of the current paper has appeared in Makadia et al. (2005). In this paper, we will present a complete theoretical and experimental treatment of our approach. In the next subsection we will discuss related approaches. Then we will motivate the Radon transform by explaining how the well-known Hough line detection can be written as a Radon integral (Deans, 1981). In Section 2 we elaborate on the spherical and rotational Fourier transforms. We extend this to incorporate the epipolar geometry and we show how to compute the Radon transform in the frequency domain. We describe the algorithm in a form that can be easily replicated and we finish with experiments.

1.1. Related Work

Structure from motion without correspondences has a history since the 80's. Most of the approaches, called *direct* motion computation, assumed a temporally dense sequence so that computation of spatio-temporal derivatives is feasible. When assuming the projection of a plane (Negahdaripour and Horn, 1987; Szeliski and Kang, 1995), the eight optical flow parameters can be estimated directly from the brightness change constraint equation. When no assumption about structure is made, several computation schemes have been proposed (Horn and Weldon, 1988). The main constraint used is depth-positiveness and usually a variational problem is solved where depth is the unknown function over the image. Direct approaches based on normal optical flow or even just its direction have been thoroughly studied by Fermuller and Aloimonos (1995) who also established formal conditions for ambiguity and instability of solutions. Jin et al. (2003) have applied a direct method for simultaneous matching of regions and 3D-motion estimation over time by exploiting photometric constraints.

Among the approaches which do not use spatio-temporal derivatives and thus can afford any amount of motion, the closest to ours are the ones by Dellaert et al. (2000), Antone and Teller (2002), and Roy and Cox (1996). In Dellaert et al. (2000), all possible assignments of 3D-points to image features are considered and the correct correspondence is established through an iterative expectation-maximization scheme where the E-step computes assignment weights and the M-step structure and motion parameters. In Antone and Teller (2002), images are already de-rotated using vanishing point correspondences and the translation is initialized via a Hough transform over all possible feature correspondences. Antone and Teller are the only ones who use the epipolar constraint and address the complexity of such a Hough transform. They propose ways to prune

the search space through feature similarity as well as limits in the parameter space. In Roy and Cox (1996), an exhaustive search in the 5D parameter space is performed where for each motion hypothesis a cost function between points in the first image and segments of the corresponding epipolar line in the second image is computed. Our approach is also related to the learning of the epipolar geometry (Wexler et al., 2003) though ours is not data-driven but requires a calibrated camera. Our approach is superior to Dellaert et al. (2000) and Antone and Teller (2002) because it is not based on an iterative process which can possibly run through all assignments. While we use an exhaustive search in parameter space, the computation of the associated "likelihood" is accomplished without iteration but directly from the spherical harmonic coefficients. Our approach is superior to Roy and Cox only in the efficient computation of each motion hypothesis. We have not described here work on motion segmentation given correspondences. The reader is referred to the application of normalizepd cuts (Shi and Malik, 1998) and the generalized PCA (Vidal and Ma, 2004) among tens of other papers on the subject. Regarding other applications of spherical harmonic analysis in computer vision, readers are referred to Basri and Jacobs (2003), Mahajan et al. (2006), and Schröder and Sweldens (1995).

2. Radon Transform

The first steps of state-of-the-art motion estimation algorithms invariably involve generating and matching features between image pairs. The assumption is that a sufficient number of these hypothesized pairs will reflect true correspondences. Any subsequent processing, such as a RANSAC motion estimation, will then terminate quickly and correctly. The problem arises when this requirement cannot be satisfied. When dealing with image pairs with small overlap, or a particularly noisy scene for feature detection, the true correspondences within a group of matched features may be very small. Our desire to process images with small overlap and to resist outliers leads us to revisit classical robust accumulation algorithms like the Hough transform. In lieu of filtering sets of image features in search of the best matches, we will treat all possible feature pairs between two images. The only discriminating measure we will consider is a similarity between features. Our signal is not an image of greyscale intensities, but rather a function which maps feature pairs to their similarities. We will accomplish our robust accumulation via a filtering which, for any camera motion, collects and *counts* all the feature pairs which satisfy a geometrical motion constraint. The counting will be weighted by the feature similarities (see Fig. 1). The filtering result provides the score for a particular

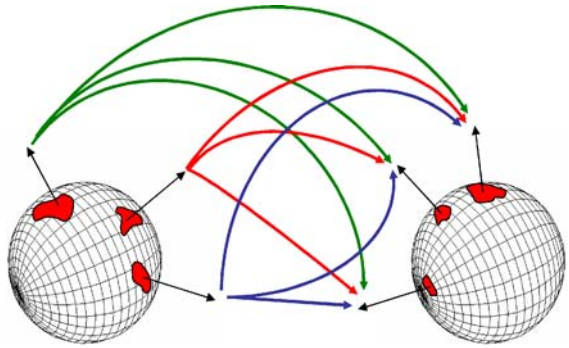

Figure 1. Concept: Instead of searching for corresponding points between images, we consider *all* feature pairs. The motion which is satisfied by the largest subset of feature pairs (weighted by a similarity measure) is considered to be the true camera motion. In the example above a weighting could be generated from the similarity between local blob structure.

motion, and in this way we can evaluate all the possible camera motions. Before presenting the concrete specification of our formulation, we introduce necessary notation and definitions which we will use throughout this section.

Consider a camera moving rigidly in space. Assuming the intrinsic calibration parameters of the camera are known (meaning we can associate with each image pixel a ray in space), we can assume that the camera model is spherical perspective. This is useful since many single-viewpoint camera systems ranging from traditional CCD cameras to fish-eye lenses and even omnidirectional cameras can be treated with a spherical projection model. In this setting, points $P \in \mathbb{R}^3$ in the world project to points on the unit sphere: $p \in \mathbb{S}^2$, where $p = P/||P||$. We will identify rigid camera motions with elements of the Euclidean motion group $SE(3)$, with one notable irregularity. Since camera translations can only be recovered up to scale, we fix the scale of the translational motion component to have unit length. Although the set of all possible camera movements can be identified with $SE(3)$, we can represent any full observable camera motion with a pair $(R, T) \in \{R \in SO(3), T \in \mathbb{R}^3, \|T\| = 1\}$. We will parameterize $SO(3)$ with ZYZ Euler angles such that $R(\alpha, \beta, \gamma) = R_z(\gamma)R_y(\beta)R_z(\alpha)$. The projection geometry in stereo pairs has been extensively studied, and it is well known that if points p and q represent projections of the same scene point in cameras separated by a motion (R, T), they must obey the coplanarity (epipolar) constraint:

$$(Rp \times q)^T T = 0 \tag{1}$$

We are now prepared to concretely develop our accumulation. As we mentioned earlier, we will not be treating an image of intensities for our robust accumulation,

but rather a function on feature pairs. We declare $g(p, q)$ to measure the similarity between points pairs in two images. Assuming an image has n pixels, the number of possible point pairs considered would be n^2, of which clearly no more than n pairs can represent true correspondences. With such a miniscule percentage of inlying point pairs, it is essential that we construct a sufficiently discriminating weighting function $g(p, q)$. In our setting it is clear a simple image-based neighborhood similarity will not suffice. Instead of using intensity information directly, we have chosen to use the popular SIFT features (Lowe, 2004), which histogram neighborhood gradient orientations. These histograms typically make up a 128-dimensional vector (which we will denote with $\tilde{p}$), which affords us many options in selecting a similarity function. For example, our weighting could depend inversely on the Euclidean distance between two feature vectors:

$$g(p, q) = e^{-||\tilde{p}-\tilde{q}||} \tag{2}$$

Alternatively, we could choose a step function:

$$g(p, q) = \begin{cases} 1 & \text{if } \|\tilde{p} - \tilde{q}\| \leq \text{Threshold} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Notice the value of $g(p, q)$ is only defined for the point pairs where we have detected features. We set $g(p, q) = 0$ whenever features were not detected at both p and q.

To perform our robust accumulation, we need a way to filter and collect all the feature pairs (p, q) from the similarity function g which satisfy the epipolar geometry given by a particular motion. To this end, we introduce the *Epipolar Delta Filter (EDF)*. The EDF has the effect of counting all the feature pairs (p, q) which satisfy the motion constraint (weighted by their feature similarities $g(p, q)$), through an inner product with g. As the EDF captures the geometry of the epipolar constraint, it must encode the possible locations of an image point p after a camera motion. We choose the most straightforward definition constructed from the epipolar constraint:

$$\Delta_{(R,T)}(p, q) = \delta((Rp \times q)^T T) \tag{4}$$

Here $\delta(x)$ is a unit impulse:

$$\delta(x) = \begin{cases} 1 & \text{if } x = 0 \\ 0 & \text{otherwise} \end{cases}$$

We can now write our robust accumulation as a filtering of a similarity function g with the EDF:

$$G(R, T) = \int_{p\in\mathbb{S}^2} \int_{q\in\mathbb{S}^2} g(p, q)\Delta_{(R,T)}(p, q)dpdq \tag{5}$$

Effectively, $G(R, T)$ is a global likelihood function as the relative likelihoods of all possible motions are computed. The correct camera motion is expected to coincide with the global peak in this grid. To generate our likelihoods, we must compute the integral (Eq. (5)) as many times as the number of samples we are considering in our discrete motion space.

If N is the number of samples in each dimension of the motion space, and M the number of features identified in each image, then the complexity of this direct approach would be on the order of $O(N^5M^2)$. This is an unacceptable load for almost any practical application. In the following sections we will demonstrate an efficient algorithm to generate the values of $G(R, T)$.

3. Motion Estimation as Correlation

In choosing to develop our global likelihood grid as a spherical filtering process, it is naturally revealed that the similarity function g is independent of the motion parameters and the EDF is independent of any feature information. For now, we will focus our attention on the EDF $\Delta_{(R,T)}$. As the direction of camera translation is the unit vector $T \in \mathbb{S}^2$, we can represent T with a rotation $R_t \in SO(3)$: $T = R_t e_3$. Here e_3 is the standard Euclidean basis vector associated with the Z axis. This allows us to parameterize the space of camera motions with a rotation pair $(R, R_t) \in SO(3) \times SO(3)$. The EDF can now be redefined as

$$\begin{aligned} \Delta_{(R,R_t)}(p, q) &= \delta\big((Rp \times q)^T R_t e_3\big) \\ &= \delta\big(\big(R_t^{-1}Rp \times R_t^{-1}q\big)^T e_3\big) \end{aligned} \tag{6}$$

If we write $R_c = R^{-1}R_t$ for the composite rotation embedding the rotational and translational terms, we see that the EDF simplifies to

$$\Delta_{(R_c,R_t)}(p, q) = \delta\big(\big(R_c^{-1}p \times R_t^{-1}q\big)^T e_3\big) \tag{7}$$

Defining the rotation operator Λ_{R_1,R_2} ($\Lambda_{R_1,R_2} f(p, q) = f(R_1^{-1}p, R_2^{-1}q)$), the EDF can be seen as just a spherical rotation of the EDF given by $(R_c, R_t) = (I, I)$:

$$\begin{aligned} \Delta_{(R_c,R_t)}(p, q) &= \delta\big(\big(R_c^{-1}p \times R_t^{-1}q\big)^T e_3\big) \\ &= \Delta_{(I,I)}\big(R_c^{-1}p, R_t^{-1}q\big) \\ &= \Lambda_{(R_c,R_t)}\Delta_{(I,I)}(p, q) \end{aligned} \tag{8}$$

We call $\Delta_{(I,I)}$ the *canonical EDF* for our parameterization. To simplify notation, we will write $\Delta(p, q)$ in place of $\Delta_{(I,I)}(p, q)$. Notice that the canonical EDF $\Delta(p, q)$ captures a translation along the Z axis and a rotation of either 0° or 180° about the Z axis. With the evolution of the EDF into Eq. (8), we can revisit our original

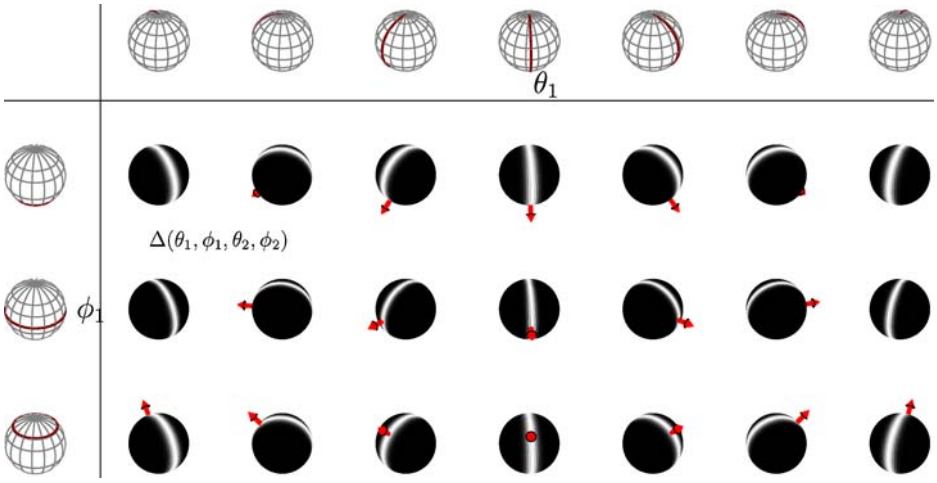

Figure 2. Here we show a 4D plot of the EDF $\Delta(\theta_1, \phi_1, \theta_2, \phi_2)$ in a 2D grid. Each plot on the sphere is a plot over (θ_2, ϕ_2) and different positions in the grid of spherical plots correspond to different choices of (θ_1, ϕ_1). The arrows (red in color) show the direction of (θ_1, ϕ_1). For the canonical EDF, corresponding to pure translation of the camera along the Z-axis, $\Delta(\theta_1, \phi_1, \theta_2, \phi_2)$ is peaked when the corresponding points are along the same longitude, i.e. when $\theta_1 = \theta_2$. Thus each arrow goes through a peak of Δ.

formulation of the global likelihood grid (Eq. (4)):

$$G(R_c, R_t) = \int_p \int_q g(p, q)\Lambda_{(R_c,R_t)}\Delta(p, q)dpdq \qquad (9)$$

This shows us that our likelihoods can be computed as a correlation between spherical functions. Figure 2 depicts the canonical EDF $\Delta(p, q)$. In the next section we will explore the theory of generalized Fourier analysis to help alleviate some of the computational burden in evaluating our likelihood function.

4. Harmonic Analysis

The spherical correlation we are considering recalls the classical signal correlations on the real line or plane. Applications of such methods include standard techniques in pattern matching. In such problems the search is for a planar shift (translational and/or rotational) which aligns a template pattern with a query image, where the location of highest correlation marks the correct alignment. These methods exploit the fact that correlations on the plane can be expressed as convolutions, and the well-known convolution theorem allows temporal convolutions to be replaced with pointwise multiplication in the spectral domain. Unfortunately, this property does not extend simply to the sphere, as convolutions and correlations on the sphere have different interpretations. Since it is not immediately clear what the relationship between the two formulations are, we will give a brief explanation. For background material, readers should consult (Helgason, 2000; Maslen and Rockmore, 1995; Sugiura, 1990).

A general definition of convolution can be given as

$$(f \star h)(x) = \int_{g\in G} f(g)h(g^{-1}x)dg$$

Here $f(x)$ and $h(x)$ are defined on some group G, and $g, x \in G$. If we take the real plane $\mathbb{R}^2$ to be a group with the action of translations, the convolution can be specifically written as

$$(f \star h)(x_1, x_2) = \int_{g_1}\int_{g_2} f(g_1, g_2)h(x_1 - g_1, x_2 - g_2)dg_1dg_2$$

This equation is the traditional form of planar convolution. Unfortunately, although the sphere is a manifold, it is not a group. We must find an alternate definition for the convolution of functions on the sphere. It is well known that the sphere is a homogeneous space of the group of 3D rotations $SO(3)$, with the isotropy subgroup of one dimensional rotations $SO(2)$ which keeps the north pole fixed (Gallier, 2005). A general definition of convolutions on homogeneous spaces can be given as

$$(f \star h)(x) = \int_{g\in G} f(g\eta)h(g^{-1}x)dg$$

Here $f(x)$ and $h(x)$ are defined on some homogeneous space of a group G, and η is given as the fixed point of the isotropy subgroup. The convolution of two functions on the sphere is given as

$$(f \star h)(x) = \int_{g\in SO(3)} f(ge_3)h(g^{-1}x)dg \; x \in \mathbb{S}^2$$

Looking closely at this definition reveals that spherical convolution betrays the traditional concept of "measuring overlap" which is implied by planar convolution. Here, points in one sphere ($f(x)$) are integrated through entire circles on the second sphere ($h(x)$). The resulting function $(f \star h)(x)$ is also defined on the sphere, hence spherical convolution reflects the properties of a filtering operator. To achieve the effect of a template matching operation, we must proceed to the general definition of correlation on homogeneous spaces:

$$c(g) = \int_x f(x)h(g^{-1}x)dx$$

As before f, h are defined on a homogeneous space of a group G, and $g \in G$ (alternatively, if we were interested in correlation on groups, we could just specify f, h to be functions on G). Identifying $\mathbb{S}^2$ as the homogeneous space of $SO(3)$ leads us to this definition of spherical correlation:

$$c(g) = \int_{x\in\mathbb{S}^2} f(x)h(g^{-1}x)dx$$

Here points on the sphere are given as unit vectors, and elements of the rotation group are given with the usual 3×3 rotation matrices. Notice that the resulting function $c(g)$ is defined not on the sphere but the group of rotations. This gives us the desired effect of measuring overlap. We rewrite this definition of spherical correlation using the notation developed earlier:

$$G(R) = \int f(\eta)\Lambda_R h(\eta)d\eta,\ f, h \in \mathcal{L}^2(\mathbb{S}^2),\quad G(R) \in \mathcal{L}^2(SO(3)) \tag{10}$$

Here $\mathcal{L}^2(\mathbb{S}^2)$ denotes square-integrability, meaning the set of functions f such that $\int |f(\eta)|^2 d\eta$ is finite. If we wish to generalize the convolution theorem to correlation on the sphere, we must be able to answer three questions: *(1) How can we compute the Fourier transform of* $f, h \in \mathcal{L}^2(\mathbb{S}^2)$ *and* $G \in \mathcal{L}^2(SO(3))$*? (2) How does the spectrum of h change under a rotation* $\Lambda_R h$*? (3) How can we compute the Fourier transform of* $G(R)$ *efficiently using the answers to questions 1 and 2?* To answer these questions we will present a minimal introduction to spherical and rotational signal processing.

4.1. *Fourier Transforms on* $\mathbb{S}^2$ *and* $SO(3)$

This treatment of spherical harmonics is based on Arfken and Weber (1966) and Driscoll and Healy (1994). In traditional Fourier analysis, periodic functions on the line (or equivalently functions on the circle $\mathbb{S}^1$), are expanded in a basis spanned by the eigenfunctions of the Laplacian. Similarly, the eigenfunctions of the spherical Laplacian provide a basis for $f(\eta) \in \mathcal{L}^2(\mathbb{S}^2)$. These eigenfunctions are the well known spherical harmonics ($Y_m^l : \mathbb{S}^2 \mapsto \mathbb{C}$), which form an eigenspace of harmonic homogeneous polynomials of dimension $2l + 1$. Consequently, the $2l + 1$ spherical harmonics for each $l \geq 0$ form an orthonormal basis for any $f(\eta) \in \mathbb{S}^2$. The $(2l + 1)$ spherical harmonics of degree l are given as

$$Y_m^l(\theta, \phi) = (-1)^m \sqrt{\frac{(2l+1)(l-m)!}{4\pi(l+m)!}} P_m^l(\cos\theta)e^{im\phi},\quad m = -l, \ldots, l \tag{11}$$

where P_m^l are the associated Legendre functions and the normalization factor is chosen to satisfy the orthogonality relation

$$\int_{\eta\in S^2} Y_m^l(\eta)Y_{m'}^{l'}(\eta)d\eta = \delta_{mm'}\delta_{ll'}, \tag{12}$$

where δ_{ab} is the Kronecker delta function. Any function $f(\eta) \in \mathcal{L}^2(\mathbb{S}^2)$ can be expanded in a basis of spherical harmonics:

$$f(\eta) = \sum_{l\in\mathbb{N}} \sum_{m=-l}^{l} \hat{f}_m^l Y_m^l(\eta) \tag{13}$$

$$\text{where } \hat{f}_m^l = \int_{\eta\in S^2} f(\eta)\overline{Y_m^l}(\eta)d\eta \tag{14}$$

The $\hat{f}_m^l$ are the coefficients of the Spherical Fourier Transform (SFT). Henceforth, we will use $\hat{f}^l$ and Y^l to annotate vectors in $\mathbb{C}^{2l+1}$ containing all coefficients or harmonics of degree l.

Using a similar approach as seen above, we can develop a Fourier transform on the rotation group $SO(3)$ (Chirikjian and Kyatkin, 2000). When considering functions $f \in \mathcal{L}^2(SO(3))$, the Fourier transform can be described as a change of basis from the group elements to the basis of irreducible matrix representations. The spherical harmonic functions Y_m^l form a complete, orthonormal set providing a basis for the representations of SO(3). Furthermore, Schur's First Lemma from fundamental representation theory shows that they also supply a basis for the irreducible representations of $SO(3)$:

$$\Lambda_R Y^l(\eta) = U^l(R)Y^l(\eta). \tag{15}$$

The matrix elements of U^l are given by

$$U_{mn}^l(R(\alpha, \beta, \gamma)) = e^{-im\gamma} P_{mn}^l(\cos(\beta))e^{-in\alpha}\quad m, n = -l, \ldots, l. \tag{16}$$

The P^l_{mn} are generalized associated Legendre polynomials which can be calculated efficiently using recurrence relations. Such an Euler angle parameterization of the irreducible representations of $SO(3)$ leads to a useful expansion of functions $f \in \mathcal{L}^2(SO(3))$:

$$f(R) = \sum_{l\in\mathbb{N}} \sum_{m=-l}^{l} \sum_{p=-l}^{l} \hat{f}^l_{mp} U^l_{mp}(R) \tag{17}$$

$$\text{where } \hat{f}^l_{mp} = \int_{R\in SO(3)} f(R)\overline{U^l_{mp}(R)}dR \tag{18}$$

The $\hat{f}^l_{mp}$, with $m, p = -l, \ldots, l$ are the $(2l+1)\times(2l+1)$ coefficients of degree l of the $SO(3)$ Fourier transform (SOFT).

Now that we have answered our first question, we can try to understand how the spectrum of a function changes under a rotation. Intuitively, we would expect a rotation to manifest itself as a modulation of the Fourier coefficients as is the case in traditional Fourier analysis. This is, in fact, the observed effect. As spherical functions are rotated by elements of the rotation group $SO(3)$, the Fourier coefficients are "modulated" by the irreducible representations of $SO(3)$:

$$f(\eta) \mapsto \Lambda_R f(\eta) \Longleftrightarrow \hat{f}^l \mapsto U^l(R)^T \hat{f}^l \tag{19}$$

The U^l matrix representations of $SO(3)$ are the spectral analogue to 3D rotations.

4.2. Rotation Estimation as Correlation

We are now prepared to address the final question regarding a generalized theorem for spherical correlation. Examining Eq. (10) more closely, we have developed the necessary tools to treat both $f(\eta)$ and $\Lambda_R h(\eta)$ with their respective Spherical Fourier expansions. Recently, (Kostelec and Rockmore, 2003; Makadia et al., 2004) have explored the computation of such a correlation in the spectral domain. Expanding the integral $\int f(\eta)\Lambda_R h(\eta)d\eta$ we have

$$G(R) = \sum_l \sum_{m=-l}^{l} \sum_n \sum_{p=-n}^{n} \sum_{k=-n}^{n} \hat{f}^l_m \overline{\hat{h}^n_p U^n_{pk}(R)} \times \int_{\eta\in\mathbb{S}^2} \overline{Y^n_k(\eta)} Y^l_m(\eta)d\eta.$$

Given the orthogonality of the spherical harmonic functions (Eq. (12)), the only nonzero terms in the summation appear when $n = l$ and $k = m$, thus

$$G(R) = \sum_l \sum_{m=-l}^{l} \sum_{p=-l}^{l} \hat{f}^l_m \overline{\hat{h}^l_p U^l_{pm}(R)}. \tag{20}$$

At this point, a direct application of the SOFT for $G(R)$ produces

$$\hat{G}^n_{qr} = \sum_l \sum_{m=-l}^{l} \sum_{p=-l}^{l} \hat{f}^l_m \overline{\hat{h}^l_p} \int_{R\in SO(3)} \overline{U^l_{pm}(R)} U^n_{qr}(R)dR$$

The orthogonality of the matrices $U^l(R)$ ($\int U^l_{mp}(R)\overline{U^n_{qr}(R)}dR = \delta_{ln}\delta_{mq}\delta_{pr}$) yields nonzero terms in the summation only when $l = n$, $m = q$, and $p = r$, resulting in this simpler expression:

$$\hat{G}^l_{mp} = \hat{f}^l_m \overline{\hat{h}^l_p} \tag{21}$$

As we had initially desired, the result of the convolution theorem can indeed be generalized to correlation on the sphere: the $SO(3)$ Fourier coefficients of the correlation of two spherical functions can be obtained directly from the multiplication of the individual SFT coefficients. In vector form, the $(2l + 1) \times (2l + 1)$ matrix of SOFT coefficients $\hat{G}^l$ is equivalent to the outer product of the coefficient vectors $\hat{f}^l$ and $\hat{h}^l$. Given $\hat{G}^l$, the inverse SOFT retrieves the desired function $G(R)$.

Recalling our original problem of filtering a feature similarity function with the Epipolar Delta Filter (Eq. (9)), we realize that we are actually correlating two functions on $\mathbb{S}^2\times\mathbb{S}^2$. As one would expect, the theory we have just introduced extends easily. The Fourier transform for any function $f \in \mathcal{L}^2(\mathbb{S}^2 \times \mathbb{S}^2)$ is given as

$$f(p, q) = \sum_{l_1} \sum_{l_2} \sum_{m_1=-l_1}^{l_1} \sum_{m_2=-l_2}^{l_2} \hat{f}^{l_1 l_2}_{m_1 m_2} Y^{l_1}_{m_1}(p) Y^{l_2}_{m_2}(q) \tag{22}$$

$$\hat{f}^{l_1 l_2}_{m_1 m_2} = \int_p \int_q f(p, q)\overline{Y^{l_1}_{m_1}(p) Y^{l_2}_{m_2}(q)}dpdq \tag{23}$$

The spectrum of $G(R_c, R_t)$ from Eq. (9) can be obtained from the Fourier transforms of g, Δ:

$$\hat{G}^{l_1 l_2}_{m_1 m_2 k_1 k_2} = \overline{\hat{f}^{l_1 l_2}_{m_1 k_1}} \hat{\Delta}^{l_1 l_2}_{m_2 k_2} \tag{24}$$

As this last equation shows, the Fourier space of our likelihood grid is six dimensional. However, we know that the space of observable motions is only five dimensional. This discrepancy arises because we identify the rotation R_t with elements of $SO(3)$ even though the translation direction is independent of the first Euler angle of rotation:

$$R_z(\alpha_1)e_3 = R_z(\alpha_2)e_3 \;\forall\; \alpha_1, \alpha_2$$

This issue is resolved easily in the following subsection.

4.3. The Canonical EDF and its Fourier Transform

The canonical Epipolar Delta Filter Δ embeds the epipolar geometry of the motions consistent with a rotation $R = I$ and translation $T = e_3$. As defined, it is only nonzero for point pairs $(p, q) \in \mathbb{S}^2 \times \mathbb{S}^2$ such that $(p \times q)^T e_3 = 0$. For any point p, the points q which satisfy this constraint must all lie on the same great circle. In particular, if we write image points with spherical coordinates θ and ϕ, then the points $p(\theta_1, \phi_1)$ and $q(\theta_2, \phi_2)$ can only satisfy the constraint $(p \times q)^T e_3 = 0$ iff $\phi_2 = \phi_1, \phi_1 + \pi$ or p or $q = \pm e_3$. Armed with this information, we can take a closer look at the Fourier transform of the EDF.

Proposition 1. *The Fourier transform of the EDF* ($\hat{\Delta}^{l_1 l_2}_{m_1 m_2}$) *is zero if and only if* l_1 *odd,* l_2 *odd,* $|m_1|$ *odd,* $|m_2|$ *odd, or* $m_1 + m_2 \neq 0$.

Proof: Let us begin by writing out the Fourier transform knowing that $\phi_2 = \phi_1, \phi_1 + \pi$:

$$\hat{\Delta}^{l_1 l_2}_{m_1 m_2} \propto \left[\int P^{l_1}_{m_1}(\cos\theta_1) \sin\theta_1 d\theta_1 \int P^{l_2}_{m_2}(\cos\theta_2) \sin\theta_2 d\theta_2 \times \int e^{i(m_1+m_2)\phi_1} d\phi_1 \right] (1 + e^{i m_1 \pi}) \tag{25}$$

Immediately we see that if $|m_1|$ is odd, then $e^{i m_1 \pi} = -1$ and the $\hat{\Delta} = 0$. Equivalently, if we had taken the expansion making a variable substitution for ϕ_1 instead of ϕ_2, we would have a trailing multiplicative term of $(1 + e^{i m_2 \pi})$, giving $\hat{\Delta} = 0$ if $|m_2|$ odd. Furthermore, the integral $\int P^{l_1}_{m_1}(\cos\theta_1) \sin\theta_1 d\theta_1 = 0$ if $m_1 = 0$ or $(l_1 + m_1)$ is odd. This means that $\hat{\Delta} = 0$ when l_1 is odd. The same argument shows $\hat{\Delta} = 0$ when l_2 is odd. The remaining integral $\int e^{-i(m_1+m_2)\phi_1} d\phi$ is only nonzero when $m_1 + m_2 = 0$, which means $\hat{\Delta} = 0$ whenever $m_1 + m_2 \neq 0$.

Now it remains to show the proposition holds in the other direction. If $\hat{\Delta} = 0$, then we know at least one of the following must be true:

1. $\int P^{l_1}_{m_1}(\cos\theta_1) \sin\theta_1 d\theta_1 = 0$
2. $\int P^{l_2}_{m_2}(\cos\theta_2) \sin\theta_2 d\theta_2 = 0$
3. $e^{-i m_1 \pi} = -1$
4. $\int e^{-i(m_1+m_2)\phi_1} d\phi_1 = 0$

The first option can only be satisfied if $l_1 + m_1$ odd or $m_1 = 0$. The second option requires $l_2 + m_2$ odd or $m_2 = 0$. The third condition requires $|m_1|$ odd (as before, we can also derive the same requirement for $|m_2|$ odd). The final option holds only if $m_1 + m_2 \neq 0$, and this completes our proof. □

We only have to consider $\hat{\Delta}^{l_1 l_2}_{m_1, -m_1}$ for $l_1, l_2, |m_1|$, even. We can now reduce the Fourier transform of the likelihood grid in Eq. (24):

$$\hat{G}^{l_1 l_2}_{m_1 m_2 k_1 - m_2} = \overline{\hat{f}^{l_1 l_2}_{m_1 k_1}} \hat{\Delta}^{l_1 l_2}_{m_2, -m_2} \tag{26}$$

Now that we have made our final simplification, we present an outline of the full algorithm in Fig. 3.

INPUT

1. A pair of spherical images I_1, I_2.

OFFLINE

1. Compute the Fourier transform $\hat{\Delta}$ of Δ from equation (23).

ONLINE

1. Detect SIFT feature sets p, q from images I_1, I_2.
2. From the cross product of the feature sets generate the similarity function g.
3. Compute the Fourier transform $\hat{g}$ of g from equation (23).
4. Generate the 5D coefficient space $\hat{G}^{l_1 l_2}_{m_1 m_2 k_1 - m_2}$ from $\hat{g}$ and $\hat{\Delta}$ as described in equation (26).
5. Using inverse Fourier transforms obtain $G(R_c, R_t)$. Note: only a partial 2D inverse transform is needed for $R_t = R(0, \beta, \gamma)$.
6. Locate (R_c, R_t) at the maxima of G.
7. Extract the correct camera motion: relative orientation between cameras is $R = R_t R_c^{-1}$, and the direction of translation is $T = R_t e_3$.

Figure 3. The full motion estimation algorithm.

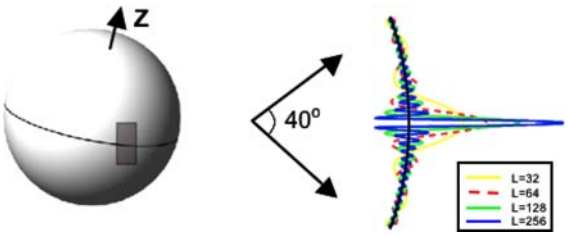

Figure 4. On the left is an spherical image of a great circle. Ideally, the function values are unity for any point on the circle, and zero otherwise. On the right is a segment of the great circle that intersects the north and south poles. The segment (which is highlighted on the left image), This shows the reconstructed function values of the delta function at the equator ($\pm 20^\circ$). The four values of the bandwidth L tested were 32, 64, 128, and 256. As L increases, the closer the approximation to an impulse, but because of the discontinuity there is also a greater overshoot (Gibbs phenomenon).

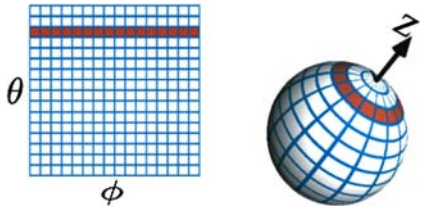

Figure 5. On the left is a grid depicting the sampling of a spherical function with bandwidth $L = 8$. Each white square is one spherical sample, and the exact location of the sample would be the middle of the square. The sampling theorem requires $2L$ uniformly spaced samples in both coordinates $\theta \in [0, \pi]$ and $\phi \in [0, 2\pi)$, hence there are 16 rows and 16 columns. The image on the right depicts the positions on the sphere of all 16^2 samples. The highlighted samples on this sphere correspond to the highlighted row of samples in the left image. One visible effect of this sampling theorem is that the sampling is dense at the north and south poles but sparse at the equator.

5. Discretization and Sampling

There are some issues we must address before we can finalize the transition from the continuous environment (integration of functions $f \in \mathcal{L}^2(\mathbb{S}^2)$) to the discrete setting (images and features). The most obvious concern relates to the Spherical Fourier Transform of a discrete spherical image. In addition to the existence of a sampling theorem, we need to be assured that the cost or complexity of the transform does not outweigh the benefits of replacing the correlation with a multiplication in the spectral domain. In other words, we require an algorithm for a discrete and fast SFT.

The bandwidth L of a spherical function f is the smallest degree such that $\hat{f}_m^l = 0, \ \forall l \geq L$. Unfortunately, the signals we are dealing with (impulse responses for the similarity function g, and great circles for the EDF), do not have a frequency limit. The bandwidth must be manually selected, and in practical terms determines how accurately we wish to approximate our function. Figure 4 shows the approximation of the EDF for different bandwidth selections. From the figure we see that even though our similarity function is represented as a sum of spherical impulses, the spectral representation is smoothed, especially for smaller values of L.

Given a function with bandwidth L, Driscoll and Healy (1994) (and later refined in Rockmore et al. (2003)), have presented a fast, discrete SFT with a sampling theorem that requires $2L$ uniformly spaced samples in each spherical coordinate (see Fig. 5). Recalling Eq. (11), a spherical harmonic is a product of a Legendre polynomial (in the longitudinal parameter θ) with a complex exponential (in the azimuthal parameter ϕ). The SFT amounts to performing many Legendre transforms in θ followed by many traditional Fourier transforms in ϕ. The more complex of the two is the Legendre transform, which can be performed fast in $O(L \log^2 L)$ (Driscoll and Healy, 1994). On the order of L Legendre transforms must be computed, which gives the total complexity of the SFT as $O(L^2 \log^2 L)$. A similar separation-of-variables approach can be applied to derive a fast and discrete $SO(3)$ Fourier transform in $O(L^3 log^2 L)$ (Kostelec and Rockmore, 2003), with a similar sampling theorem.

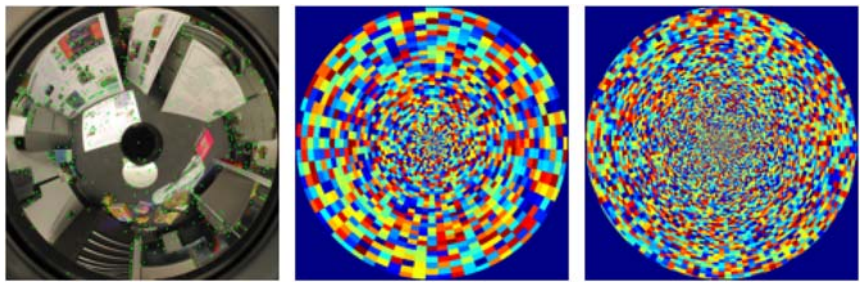

Figure 6. On the left is an image from an omnidirectional sensor, with a field of view of 212°. In the middle is a spherical image with bandwidth $L = 32$ mapped onto the omnidirectional image plane. Each segment in this image corresponds to one pixel in the spherical image. This shows the quantization or binning effect seen when mapping points from a high-res image to a low-bandwidth spherical function. On the right is the same effect for a bandwidth $L = 40$.

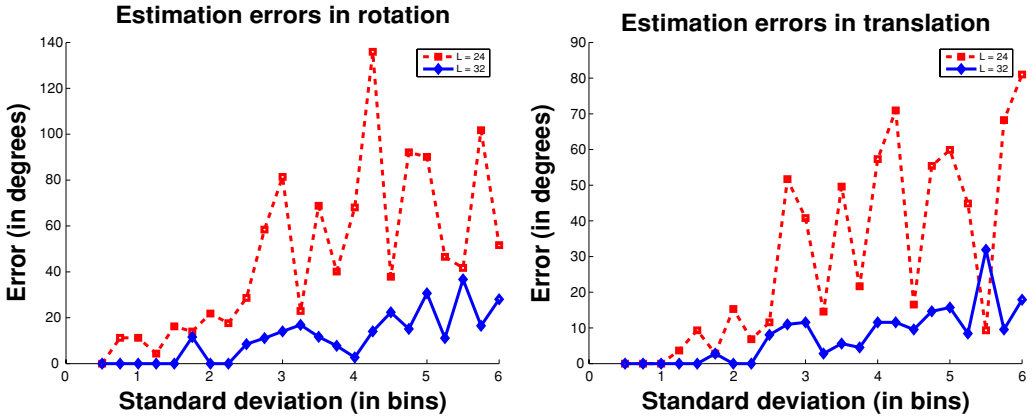

Figure 7. Results of a simulation testing the robust accumulation in the presence of Gaussian noise. The locations of spherical point correspondences are perturbed with Gaussian noise in each spherical coordinate. The standard deviation of the noise distribution is given here in pixels, and the error is computed by measuring the distance of the estimated solution from the correct solution in the 5D motion space. On the left is the error in the estimated rotation (the angular distance between two rotation matrices is computed as $\arccos((trace(R_1^{-1}R_2)-1)/2)$), and on the right is the error in baseline direction. The dashed plots (in red) represent the simulation performed with bandwidth $L = 24$. In this case, a standard deviation of one unit corresponds to 7.5° and 3.8° in the spherical coordinates ϕ and θ. The solid plots (in blue) are for $L = 32$, where a standard deviation of one pixel corresponds to 5.6° and 2.8° in the spherical coordinates. For this higher bandwidth, the results are still accurate in presence of significant noise.

Recall from Eq. (9) that we are parameterizing our motion space with rotations, which in turn are parameterized with ZYZ Euler angles. Let us use the angles α, β, γ, θ, and ϕ to denote each of the five dimensions of our motion space, so that $R_c = R(\alpha, \beta, \gamma)$, $R_t = R(0, \theta, \phi)$, and $\alpha, \gamma, \phi \in [0, 2\pi)$, $\beta, \theta \in [0, \pi]$. If we fix L as the bandwidth of our similarity function g and EDF Δ, and we follow the algorithm in Fig. 3 using the SFT and SOFT routines detailed in Rockmore et al. (2003) and Kostelec and Rockmore (2003), the angles α, γ, and ϕ will be sampled at

$$\alpha_j, \gamma_j, \phi_j = \frac{\pi j}{L}, j = 0, 1, \ldots, 2L-1 \qquad (27)$$

The angles β, and θ will be sampled at

$$\beta_k, \theta_k = \frac{\pi(2k+1)}{4L}, k = 0, 1, \ldots, 2L-1 \qquad (28)$$

The total number of samples in G is thus $32L^5$. In practice, this forces us to select lower values for L, such as 32. Although we are capturing high resolution images and locating image features with sub-pixel accuracy, the effective resolution of one spherical image is just $2L \times 2L$. The experiment detailed in Fig. 7 shows just how our algorithm reacts when the feature locations are affected by Gaussian noise when using such "low-resolution" spherical images. Figure 6 shows the relationship between the uniform angular spacing of the spherical samples and the original image domains of different single-viewpoint cameras. It is clear for small L many pixels from a high-res perspective or omnidirectional image will map to the same spherical sample, and since we will detect features on the original images we must clarify how to generate a discrete version of our similarity function g.

The sampling theorem requires $2L$ samples in each angle, which means every spherical function must have $4L^2$ samples, and g must therefore have $16L^4$ samples. Let us write (p^j, q^k), $j, k = 1, 2, \ldots, 4L^2$ for the samples of g. Assume we are given two input images I_1, I_2

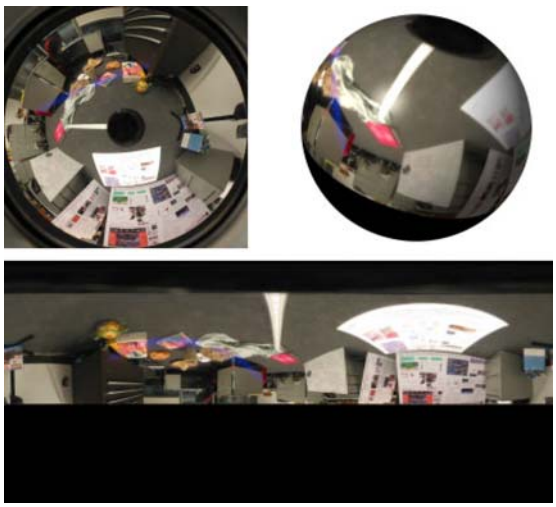

Figure 8. Top Left: a parabolic catadioptric image. Bottom: the corresponding image on a uniformly sampled spherical grid. As the parabolic mirror images only a little more than half the sphere, you can see the lower portion of the spherical image contains no information. Top Right: the spherical image as it would appear on the surface of the sphere.

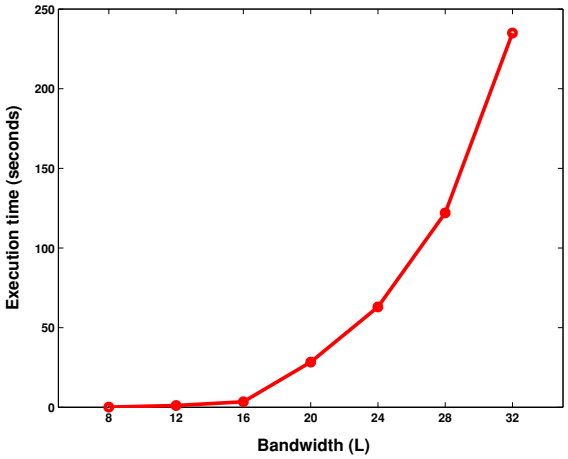

Figure 9. Timings of our algorithm for various bandwidth choices. The execution times are for step 3 through step 7 (see Fig. 3).

on which we detect N_1 and N_2 features, respectively. We denote Q as the set of all possible feature pairs (note that Q has $N_1 N_2$ elements), and each element of Q has an associated weight given by Eq. (2) (or Eq. (3)). The value of the discrete similarity function at a sample (p^j, q^k) is just the sum of the weights of all elements of Q that have this sample (p^j, q^k) as the nearest neighbor. This process has the effect of just quantizing the continuous similarity function. Whenever different point pairs are quantized into the same discrete sample, their similarity weights are simply combined.

6. Experiments

In this section we will present the results of the motion estimation algorithm on real image sequences. We begin by describing the spherical camera system which we use for our experiments.

6.1. *Spherical Image Acquisition*

One of the benefits of choosing to model our camera with a spherical perspective projection is that it enables us to unite a number of single-viewpoint camera systems. Our experiments were performed with a catadioptric camera system along with a traditional digital camera.

The projection model of a central catadioptric system is equivalent to a spherical projection followed by a

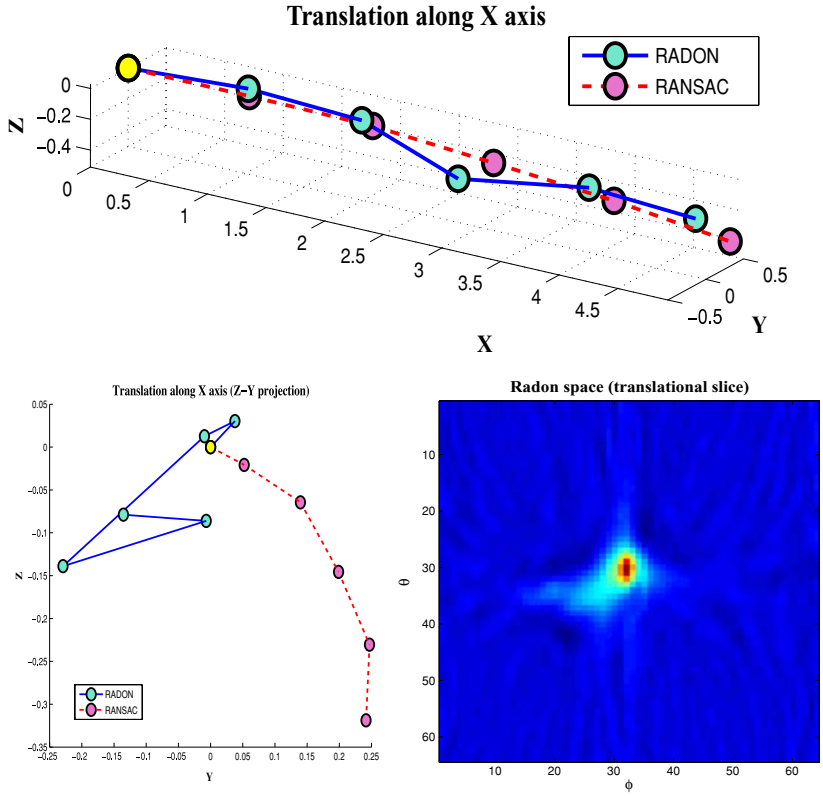

Figure 10. Top: the estimated trajectory of the camera. In solid blue (light) is the Radon estimation, in dashed red (dark) is the RANSAC computation, and the yellow circle marks the starting position. Bottom Left: A projection of the trajectory onto the $Z - Y$ plane showing the deviation of the estimated positions from the X axis. Bottom Right: the R_t slice of the grid G where the maxima was found.

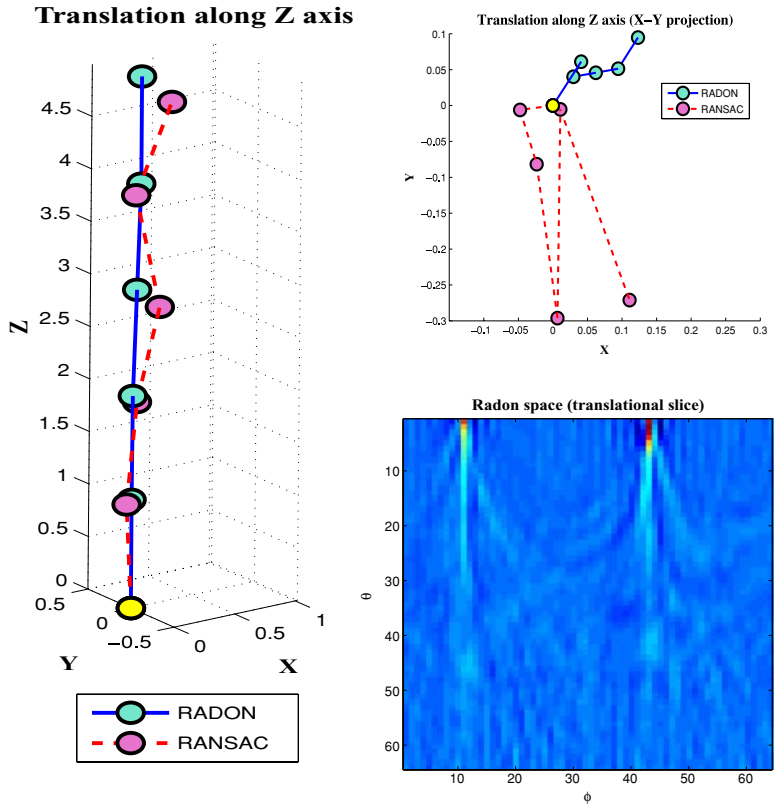

Figure 11. Left: the estimated trajectory of the camera. In solid blue (light) is the Radon estimation, in dashed red (dark) is the RANSAC computation, and the yellow circle marks the starting position. Top Right: A projection of the trajectory onto the $X-Y$ plane showing the deviation of the estimated positions from the Z axis. Bottom Right: the R_t slice of the grid G where the maxima was found (notice the peak is locate at $\theta \approx 0$, which corresponds to the correct translation along Z).

projection onto the plane (Geyer and Daniilidis, 2001). If calibrated, such a sensor enables us to interpolate spherical perspective images. Our system consisted of a Canon Powershot G2 digital camera fastened to a parabolic mirror attachment from RemoteReality^{TM} (Nayar, 1997). Being that the mirror's field-of-view is 212°, the camera captures slightly more than a hemisphere of information. Figure 8 shows a sample catadioptric image obtained from a parabolic mirror and its corresponding projection onto the sphere.

6.2. *Results*

We proceed to show experimental results of our algorithm tested on a sequence of real omnidirectional images. The running time of our algorithm for various bandwidth choices is shown in Fig. 9. For our tests, we assumed a function bandwidth of $L = 32$, which left us with a spatial resolution of $2L = 64$ samples in each of the five dimensions of our motion space. For comparison, we employed RANSAC to estimate the essential matrix. Although it seems natural to use RANSAC in the presence of outliers, there are two crucial issues which would prevent a naive implementation from being operative. First is the volume of outliers. Assuming the number of features detected in each of two images is N, there are N^2 possible feature pairs of which at most N are inliers. Since the inlier rate is no more than $1/N$ (for a typical scenario with $N = 1000$, the inlier rate is at most 0.1%), the likelihood of selecting a minimal set of true correspondences is negligible. To this end, we discarded all but the best matching pairs during the random sampling stage. We retained only approximately 0.025% of the possible feature pairs (e.g. this translates to 250 feature pairs from a set of 10^6 possible pairs). The second issue is in determining the termination threshold of the RANSAC algorithm. In order to perform a proper evaluation of our algorithm, we implemented a best-case RANSAC which does not have a termination threshold but rather iterates 50, 000 times. The essential matrix which satisfies the most feature pairs (weighted with $g(p, q)$) is selected as the motion. This ensures that a manual selection of the termination threshold may not be set too low to allow termination for an inferior motion. We have evaluated our Radon estimation alongside this modified RANSAC in order to provide an

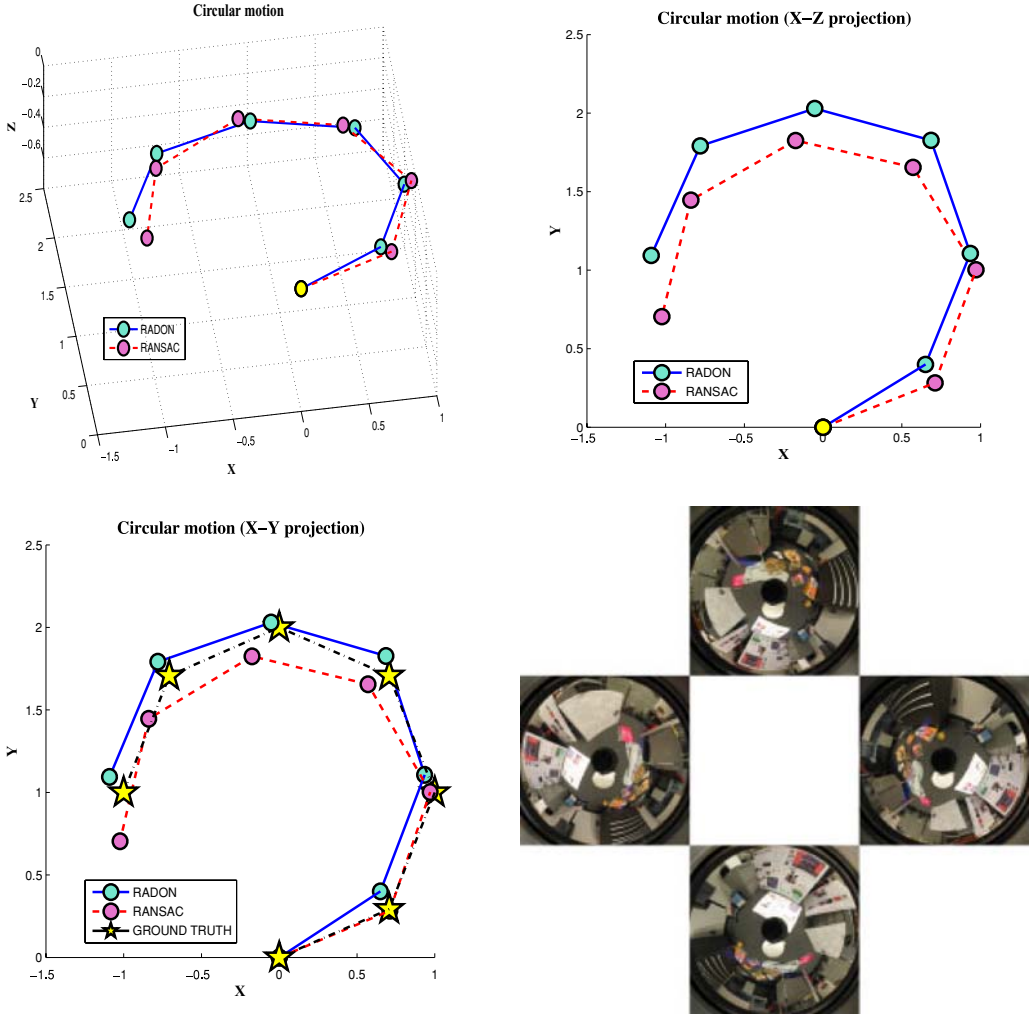

Figure 12. A camera moving along a circular path. Top left: In solid blue (light) is the Radon estimation, in dashed red (dark) is the RANSAC. Top right: A projection onto the $X - Z$ plane showing the deviation from the plane of the turntable. Bottom left: An overhead view. The yellow stars are the observed ground truth positions of the camera. Bottom right: four images from the sequence. Even though the dominant motion is rotation, the translation is still effectively detected by the Radon.

alternate method which is comparable to ours. In some of the following experimental results, the RANSAC performs very well and this is only because we have tuned these parameters quite finely. The similarity function in Eq. (2) was used for the experiments depicted in Fig. 10 through Fig. 12, while Eq. (3) was used for the remainder.

We begin with a pure translational sequence of images. By fixing and sliding our camera along a rigid beam, we were able to generate two sequences of translational motion along the X and Z axes of the camera frame. Fixing the magnitude of motion between each frame, we were able to plot the estimated camera trajectory in Fig. 10. In general, there are four possible rotation and translation pairs which will satisfy a particular epipolar constraint. These solutions correspond to the true solution, a baseline reversal, a camera rotation of 180° about the baseline (commonly referred to as the "twisted pair" configuration), or a twisted pair with baseline reversal. If the true motion is given by (R, T), the other three motions which satisfy the same epipolar constraint are given by $(R, -T)$, $(e^{\hat{T}\pi}R, T)$, and $(e^{\hat{T}\pi}R, -T)$ (note that $e^{\hat{T}\pi}$ gives a rotation of 180° about the T axis). In order to identify the expected locations of the four peaks in our likelihood space for the correct motion, we must remember that we identify elements of this five dimensional motion space with the pair (R_c, R_t) where $R = R_t R_c^{-1}, T = R_t e_3$. If we define $R'_t = R_z(\gamma)R_y(\beta)$ so that $-T = R'_t e_3$, then we can expect the four peaks to be located at $(R^{-1}R_t, R_t)$, $(R^{-1}R'_t, R'_t)$, $((e^{\hat{T}\pi}R)^{-1}R_t, R_t)$, and $((e^{\hat{T}\pi}R)^{-1}R'_t, R'_t)$. In the figures, for example, when we show a 2D translational slice

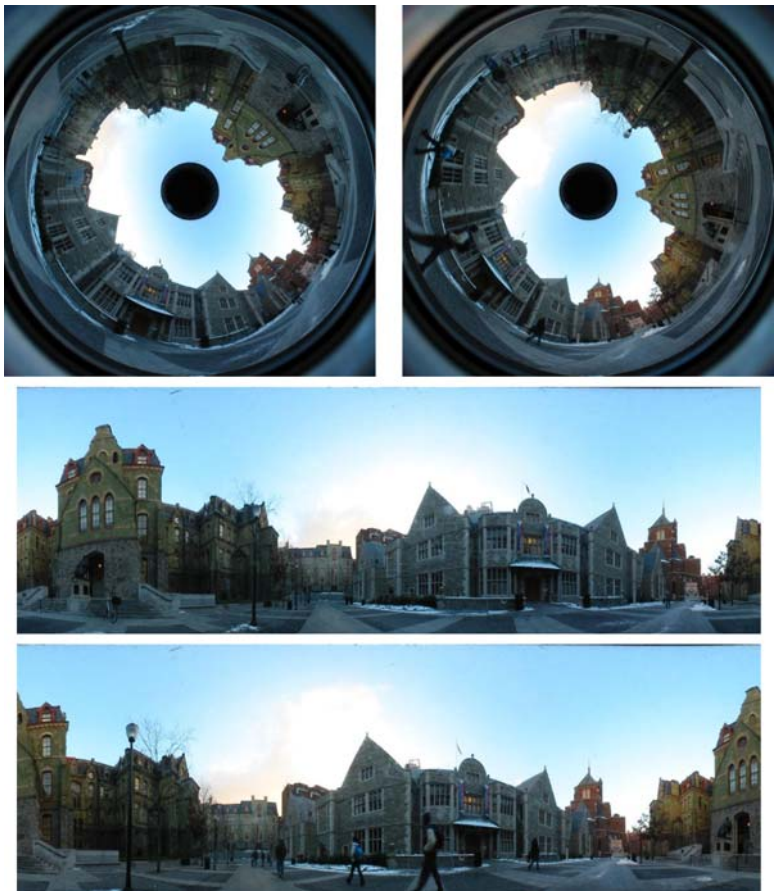

Figure 13. Top row: two representative images from a sequence of outdoor images. The motion between image positions is over five meters. At each position, the equatorial plane of the spherical image is roughly aligned to be parallel with the ground plane to provide a rough, partial ground truth of the motion. The image sequence also contains some dynamic scene content as there were people moving throughout the scene as the images were taken. The bottom two images are the spherical projections of the original omni images. Only the visible band on the sphere is shown here.

with a peak at R'_t, this slice can be generated from the bins corresponding to the rotation $R^{-1}R'_t$. In Fig. 10, the slice shown depicts a peak at $R_t(0, \frac{\pi}{2}, \pi)e_3 = -e_1$.

A similar experiment was performed with the camera moving along the Z axis. The motion was recovered from pairs of consecutive images, with the estimated camera path shown in Fig. 11. Our Radon estimation has a smaller deviation from the observed ground truth Z axis than the RANSAC estimation.

In order to test both rotations and translations while recording ground-truth observations, we positioned the camera at the outside edge of a turntable. This allowed us to capture images from the camera moving around in a circle. There was a 45° rotation between each of the images in this sequence, and the estimated camera positions are shown in Fig. 12. Although the Radon's trajectory estimate deviates slightly from the plane, the positions as seen from the overhead view coincide with the recorded ground truth more accurately than the RANSAC estimation. After 6 pairwise tests, there was little error accumulation in estimating the trajectory.

We now discuss results of an experiment from a sequence of images from an outdoor environment. Figure 13 shows a representative selection of images from this sequence. Figure 14 shows some results from the motion estimation. Epipolar lines are drawn to allow visual confirmation of the method's accuracy. Figure 15 displays the obtained camera trajectory using the visualization tools provided in the Epipolar Geometry Toolbox (Mariottini and Prattichizzo, 2005). This same trajectory is projected onto the $X - Z$ plane in Fig. 16 to show the deviation from the ground plane which is known to be (approximately) the correct plane of motion. This planar motion also restricts the axis of rotation to align with the Z axis, and Fig. 17 shows just how closely the measured rotations reflect this property.

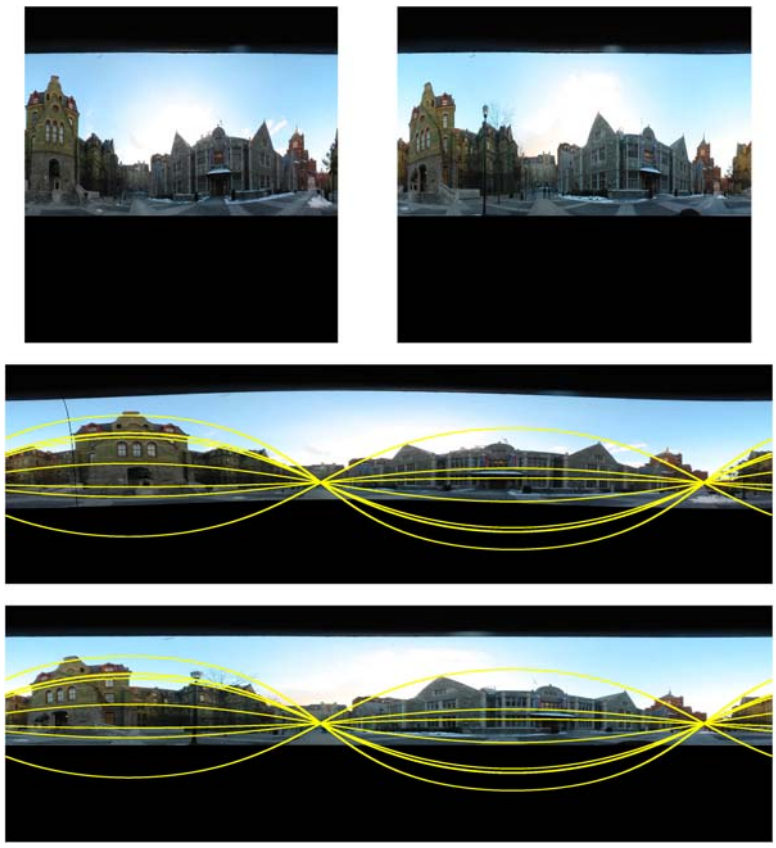

Figure 14. On the top row are a pair of images from a sequence for which the motion was estimated. The bottom two rows show the images after they have been rotationally aligned. Epipolar circles have been overlaid onto the images. Since the images have been rotationally aligned, points which lie along these circles in one image will lie along the same circle in the second image. The intersection of these circles mark the focus of expansion and contraction, which define the direction of translation between this image pair. The rotation between image pairs was estimated at approximately 45°, and as the focus of expansion shows the translation was roughly in the equatorial plane.

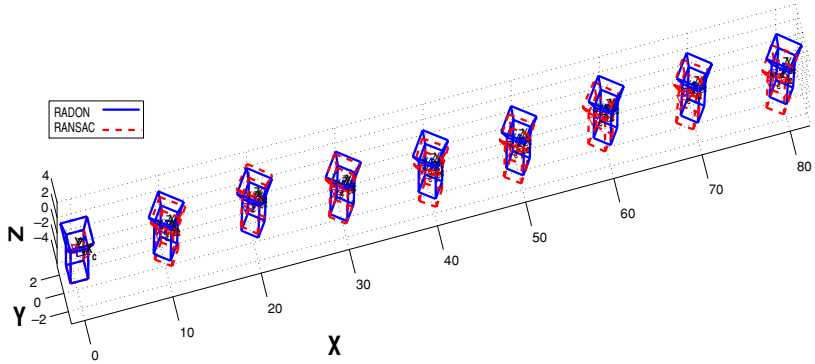

Figure 15. This figure shows the camera trajectory estimated from a sequence of images (see Fig. 13 for sample images). The camera frames drawn with solid (blue) lines depict the trajectory estimated using the Radon transform, while the dashed (red) lines show the RANSAC trajectory. In this sequence the motion is known to be approximately planar in the equatorial plane. Since the magnitude of camera motion cannot be recovered from pairs of images alone, we have fixed the distance between camera positions to be 10 units for visual purposes.

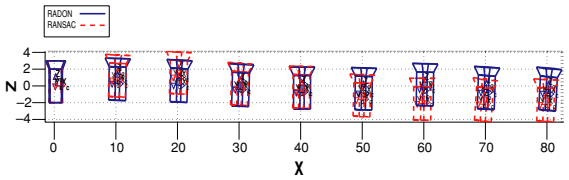

Figure 16. This figure shows the camera trajectory estimated from a sequence of images (see Fig. 13 for sample images), projected onto the ZX plane. The camera frames drawn with solid (blue) lines depict the trajectory estimated using the Radon transform, while the dashed (red) lines show the RANSAC trajectory. The motion is known to be planar (on the equatorial plane) and the Radon estimate reflects this more accurately. Since the magnitude of camera motion cannot be recovered from images alone, we have set the distance between camera positions to be 10 units for visual purposes.

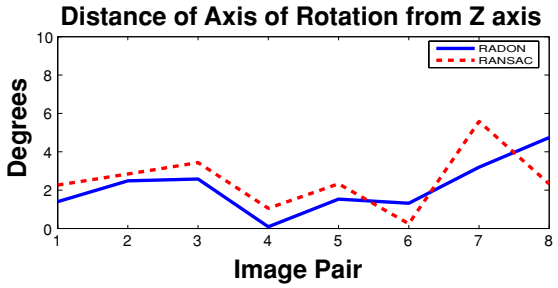

Figure 17. As the camera motion in this sequence (see Fig. 13 for sample images) is known to be roughly planar, we know that the axis of rotation must align with the Z axis. This plot shows for all eight image pairs in the sequence the distance in degrees of the estimated axis of rotation from the Z axis. The solid line (blue) is the estimate from our Radon integral, and the dashed line (red) is the RANSAC estimate.

7. Conclusion

We have presented a novel approach for the computation of 3D-motion from two views without correspondences. It is based on the generation of a global likelihood function on the space of all observable camera motions. Given today's computing power, it is not the search through this likelihood function but rather the combinatorial explosion of all possible correspondences that is intractable. Instead of traversing all possible correspondence assignments, our method computes for each motion hypothesis a correlation function which considers only feature pairs satisfying the epipolar constraint. Such a formulation can be expressed as a correlation integral if the integration path can be written as a group action over the domain of integration. In this case, the integral can be computed as an inner-product in the Fourier domain. The bandwidth limitation affects directly the resolution of the parameter space and it is indeed our future work to establish a "space localization" using wavelets. Such a localization in the parameter space would also allow a constrained search when prior distributions of motion are established causally through time. In that case, we could also achieve near real-time performance which right now is impossible in all correspondence-free approaches.

References

Antone, M. and Teller, S. 2002. Scalable, extrinsic calibration of omni-directional image networks. *International Journal of Computer Vision*, 49:143–174.

Arfken, G. and Weber, H. 1966. *Mathematical Methods for Physicists*. Academic Press.

Basri, R. and Jacobs, D.W. 2003. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):218–233.

Chirikjian, G. and Kyatkin, A. 2000. *Engineering Applications of Noncommutative Harmonic Analysis: With Emphasis on Rotation and Motion Groups*. CRC Press.

Daniilidis, K. and Spetsakis, M. 1996. Understanding noise sensitivity in structure from motion. In *Visual Navigation*, Y. Aloimonos (Ed.). Lawrence Erlbaum Associates: Hillsdale, NJ, pp. 61–88.

Deans, S. 1981. Hough transform from the radon transform. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 3:185–188.

Dellaert, F., Seitz, S., Thorpe, C., and Thrun, S. 2000. Structure from motion without correspondence. In *CVPR*. Hilton Head Island, SC.

Driscoll, J. and Healy, D. 1994. Computing fourier transforms and convolutions on the 2-sphere. *Advances in Applied Mathematics*, 15:202–250.

Fermuller, C. and Aloimonos, J. 1995. Direct perception of three-dimensional motion from patterns of visual motion. *Science*, 270:1973–1976.

Gallier, J. 2005. Notes on group actions, manifolds, lie groups, and lie algebras. `http://www.cis.upenn.edu/~cis610/lie1.pdf/`.

Geyer, C. and Daniilidis, K. 2001. Catadioptric projective geometry. *International Journal of Computer Vision*, 43:223–243.

Geyer, C., Sastry, S., and Bajcsy, R. 2004. Euclid meets fourier. In *Workshop on Omnidirectional Vision*. Prague.

Helgason, S. 2000. *Groups and Geometric Analysis: Integral Geometry, Invariant Differential Operators, and Spherical Functions*. American Mathematical Society.

Horn, B. and Weldon, E. 1988. Direct methods for recovering motion. *International Journal of Computer Vision*, 2:51–76.

Jin, H., Favaro, P., and Soatto, S. 2003. A semi-direct approach to structure from motion. *The Visual Computer*, 19:1–18.

D.H. Jr., Rockmore, D., Kostelec, P., and Moore, S. 2003. FFTs for the 2-sphere—improvements and variations. *The Journal of Fourier Analysis and Applications*, 9(4):341–385.

Kostelec, P.J. and Rockmore, D.N. 2003. FFTs on the rotation group. In *Working Paper Series.* Santa Fe Institute.

Lowe, D. 2004. Sift (scale invariant feature transform): Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110.

Mahajan, D., Ramamoorthi, R., and Curless, B. 2006. A theory of spherical harmonic identities for BRDF/lighting transfer and image consistency. In *European Conference on Computer Vision*, vol. IV, pp. 41–55.

Makadia, A., Geyer, C., Sastry, S., and Daniilidis, K. 2005. Radon-based structure from motion without correspondences. In *IEEE Conf. Computer Vision and Pattern Recognition*, San Diego.

Makadia, A., Sorgi, L., and Daniilidis, K. 2004. Rotation estimation from spherical images. In *Proc. Int. Conf. on Pattern Recognition*. Cambridge, UK.

Mariottini, G. and Prattichizzo, D. 2005. EGT: A toolbox for multiple view geometry and visual servoing. *IEEE Robotics and Automation Magazine*, 3(12).

Maslen, D. and Rockmore, D. 1995. Generalized FFTs—a survey of some recent results. In *Proceedings of the DIMACS Workshop on Groups and Computation*.

Nayar, S. 1997. Catadioptric omnidirectional camera. In *IEEE Conf. Computer Vision and Pattern Recognition*. Puerto Rico, pp. 482–488.

Negahdaripour, S. and Horn, B. 1987. Direct passive navigation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 9:168–176.

Oliensis, J. 2000. A critique of structure from motion algorithms. *Computer Vision and Image Understanding*, 80:172–214.

Roy, S. and Cox, I. 1996. Motion without structure. In *Proc. Int. Conf. on Pattern Recognition*. Vienna, Austria.

Schröder, P. and Sweldens, W. 1995. Spherical wavelets: Efficiently representing functions on the sphere. In *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*. ACM Press: New York. pp. 161–172.

Shi, J. and Malik, J. 1998. Motion segmentation and tracking using normalized cuts. In *Proc. Int. Conf. on Computer Vision*.

Sugiura, M. 1990. *Unitary Representations and Harmonic Analysis: An Introduction*. second edition. North Holland, Amsterdam.

Szeliski, R. and Kang, S.B. 1995. Direct methods for visual scene reconstruction. In *IEEE Workshop on Representations of Visual Scenes*, pp. 26–33.

Vidal, R. and Ma, Y. 2004. A unified algebraic approach to 2-D and 3-D motion segmentation. In *European Conference on Computer Vision*, pp. 1–15.

Wexler, Y., Fitzgibbon, A., and Zisserman, A. 2003. Learning epipolar geometry from image sequences. In *IEEE Conf. Computer Vision and Pattern Recognition*. Wisconsin.