Radon-based Structure from Motion Without Correspondences

Ameesh Makadia* Christopher Geyer† Shankar Sastry† Kostas Daniilidis*

*University of Pennsylvania
Philadelphia, PA 19104
{makadia, kostas}@cis.upenn.edu

†University of California, Berkeley
Berkeley, CA 94720
{cgeyer, sastry}@eecs.berkeley.edu

Abstract

We present a novel approach for the estimation of 3D-motion directly from two images using the Radon transform. We assume a similarity function defined on the cross-product of two images which assigns a weight to all feature pairs. This similarity function is integrated over all feature pairs that satisfy the epipolar constraint. This integration is equivalent to filtering the similarity function with a Dirac function embedding the epipolar constraint. The result of this convolution is a function of the five unknown motion parameters with maxima at the positions of compatible rigid motions.

The breakthrough is in the realization that the Radon transform is a filtering operator: If we assume that images are defined on spheres and the epipolar constraint is a group action of two rotations on two spheres, then the Radon transform is a convolution/correlation integral. We propose a new algorithm to compute this integral from the spherical harmonics of the similarity and Dirac functions. The resulting resolution in the motion space depends on the bandwidth we keep from the spherical transform. The strength of the algorithm is in avoiding a commitment to correspondences, thus being robust to erroneous feature detection, outliers, and multiple motions. The algorithm has been tested in sequences of real omnidirectional images and it outperforms correspondence-based structure from 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 problem of data association and estimation: to estimate 3D motion we must consider only correspondences induced by that motion, but to segment we must know the correspondences. Outlier rejection and independent motion segmentation 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 many outliers and multiple motions. It 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 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 such a 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. While we are familiar with convolution as an inner product with a shifted kernel, here it is not obvious what

[*] The authors are grateful for support through the following grants: NSF-IIS-0121293, NSF-EIA-0324977, NSF-CNS-0423891, NSF-IIS-0431070, and ARO/MURI DAAD19-02-1-0383.

[†] The author is grateful for the support through the following grants: DAAD-19-02-1-0383, Boeing sub-contract Z40705R of DARPA funded SEC program managed by AFRL, and NSF-IIS-0122599.

the domain is and what is shifted. Abstract harmonic analysis tells us that convolutions can be generalized to other domains on which groups (similar to shifts) act. In our case the domain is the cross-product of two rotated spheres and we will show that the acting group is a cross-product of rotations. After applying a modulation-like theorem to the spherical Fourier transform, 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 harmonic coefficients preserved determines the resolution of the motion space. Obviously, the approach can work on arbitrarily large motions.

We have built an end-to-end system, from images to motion parameters. We extracted hundreds of SIFT features [11] for which we defined their similarity function proportional to the Euclidean norm of the attribute vectors and we computed the spherical harmonics of the similarity function as the input to the correlation integral. The only threshold of the approach is the cut-off frequency of the harmonic coefficients which determines the resolution of the motion space. This “low-pass” operation has the appealing property of quantizing the motion space and allowing rough but faster estimates. In the experiments, we use as input hemispherical omnidirectional images. We should point out to the reader that this is not an omnidirectional structure from motion approach. 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 [14, 2]. 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 transform can be written as a convolution/correlation integral which can be computed from the spherical harmonic coefficients of the image similarity function.
- In real experiments, we compare our algorithm to a RANSAC-based approach in the presence of hundreds of outliers. In simulated experiments, we show how multiple motions are detected as maxima of the strength function in motion space.

The approach paves the way for several other motion estimation problems where the constraints can be written as convolution kernels. Currently, the main drawback is the computation time which allows the algorithm to be applied only “after action.”

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 [3]. In section 2 we elaborate on the Radon transform which is known in harmonic analysis to be written as a convolution. 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 [13, 17], 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 [8]. 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 et al. [6] who also established formal conditions for ambiguity and instability of solutions. Jin et al. [9] 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 spatiotemporal derivatives and thus can afford any amount of motion, the closest to ours is the ones by Dellaert et al. [4], Antone and Teller [1], and Roy and Cox [15]. In [4], 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 [1], 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 [15], 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 [19] though ours is not data-driven but requires a calibrated camera. Our approach is superior to [4] and [1] 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 normalized cuts [16] and the generalized PCA [18] among tens of other papers on the subject.

2 Radon transform

We begin with an introduction to the traditional Hough transform as it applies to finding lines in images. In this setting the data points are image pixels and the discrete parameter space is a set of lines. Conceptually, for each image pixel, the Hough transform contributes a vote to all the lines it lies along. This vote is weighted by the likelihood that the point under consideration is indeed an edge pixel (e.g. the gradient magnitude). Equivalently, we could describe this computation as a traversal through the parameter space instead of the data space. The vote total for each line can be generated by counting the number of image pixels the line goes through, weighted by the likelihood that each pixel is an edge pixel. In the continuous case, this computation could be written as the following integral

$$G(\rho,\theta)=\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\delta(\rho-x\cos\theta-y\sin\theta)dxdy$$

Here $g(x,y)$ is a weighting function which could store the gradient lengths of each pixel and δ is a soft characteristic function which measures how close the edge pixel (x,y) is to the line given by (ρ,θ). This integral transformation from data space to parameter space is often referred to as the Radon transform. We would like to use similar intuition to formulate a transform which will identify the unknown motion parameters.

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 projection. 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 this spherical projection model. In this setting, points $P\in\mathbb{R}^3$ in the world project to image points $p\in\mathbb{S}^2$, where $p=P/||P||$. If a camera undergoes a rigid motion described by $(R,T)\in SE(3)(R\in SO(3),T\in\mathbb{R}^3)$, it is well known that the projections p and q obey the epipolar constraint:

$$(Rp\times q)^T t=0 \qquad (1)$$

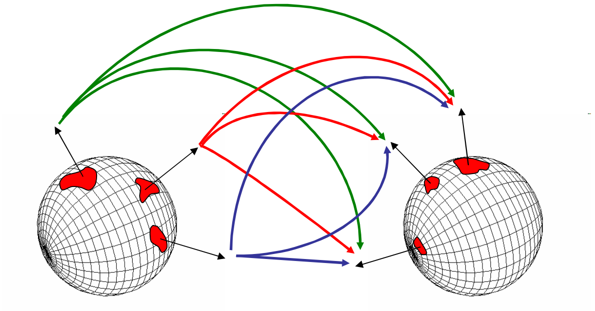

Figure 1. Concept: Instead of searching for corresponding points between images, we consider *all* feature pairs. The motion which is satisfied by the largest subset of feature pairs (weighted by a similarity measure) is considered to be the true camera motion. In the example above a weighting could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

If we were to follow the blueprint of the integral transform described earlier, we would define our parameter space to be the group of all possible rigid camera motions and our data space to be the set of all point pairs between two images. Our integral transform would look like

$$G(R,t)=\int_{p\in\mathbb{S}^2}\int_{q\in\mathbb{S}^2} g(p,q)\Delta(Rp,q,t)dpdq \qquad (2)$$

Here the soft characteristic function $\Delta(Rp,q,t)=\delta(Rp\times q)^T\, t)$, measures how close the feature pair (p,q) comes to satisfying the motion constraint (1), and $g(p,q)$ is a measure of how likely the points p,q are the projections of the same scene point. For each motion given by (R,t), the integral (2) counts the number of point pairs which satisfy the motion constraint, weighted by the likelihood that the point pair represents the same scene point (see figure 1). Take a moment to imagine a discretized evaluation of Radon the integral. 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 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 perform feature extraction in the image. Thus, instead of considering all the pixel locations in an image, we only use the positions where features can been detected. We have chosen to use the popular SIFT features [11], which histogram neighborhood gradient orientations at peaks and valleys of difference-of-gaussians. These histograms typically make up a 128-dimensional vector, which allows us to create a very simple weighting function based on the Euclidean distance between two such vectors:

$$g(p,q) \quad = \quad e^{-||p-q||_2} \qquad (3)$$

The two functions g, Δ have now been concretely defined. We could generate a solution to the ego-motion problem by computing $G(R,t)$ directly. Computationally, if we assume the number of samples in each dimension of our parameter space is N, and the number of features identified in each image is M, 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. For a rigorous look at the combinatorics of this problem, see the Appendix of [1]. In the following sections we will demonstrate an efficient algorithm to generate the values of $G(R,t)$.

3 Motion estimation as correlation

A cursory glance at our formulation of $G(R,t)$ reveals $g(p,q)$ is independent of the motion. Thus we can focus our attention on Δ. So far we have identified camera motions with an $R \in SO(3)$, and a unit vector t. Since $t \in \mathbb{S}^2$, we can represent t with a rotation so that $t = R_t e_3$, where 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 $(R, R_t) \in SO(3) \times SO(3)$. $\Delta(Rp, q, t)$ can now be written as

$$\begin{aligned}\Delta(Rp, q, R_t) &= \delta((Rp \times q)^T R_t e_3) \\ &= \delta((R_t^T Rp \times R_t^T q)^T e_3) \qquad (4)\end{aligned}$$

We will write $R_c = R^T R_t$ for the composite rotation embedding the rotational and translational terms. We have conveniently written Δ in the form of (4) to highlight it as a function defined on the space $\mathbb{S}^2 \times \mathbb{S}^2$:

$$\Delta(R_c^T p, R_t^T q) = \delta((R_c^T p \times R_t^T q)^T e_3)$$

In this setting, the canonical camera motion, defined by $R_c = R_t = I$, is represented by $\Delta(p,q) = ((p \times q)^T e_3)$, which represents a translation along the Z axis and a rotation of either 0° or 180° about the Z axis.

Define the rotation of spherical functions with the operator $\Lambda_{(R_1,R_2)} f(p,q) \equiv f(R_1^T p, R_2^T q)$. We see that the Δ for any camera motion (R_c, R_t) can be generated from the rotation of the canonical Δ : $\Lambda_{(R_c,R_t)}\Delta(p,q) = \Delta(R_c^T p, R_t^T q)$. Revisiting our transform (2), we can write

$$G(R_c, R_t) \;=\; \int_p \int_q g(p,q)\Lambda_{(R_c,R_t)}\Delta(p,q)dpdq \quad (5)$$

Instead of recomputing Δ for every motion, we only need to understand how the canonical Δ *rotates*. In the following section, we will use this crucial fact to explore a spectral correlation technique which will enable us to compute $G(R_c, R_t)$ directly without traversing the space of all possible camera motions.

4 Harmonic analysis

The inner product computed in (5) measures the correlation between two functions $g, \Delta \in \mathcal{L}^2(\mathbb{S}^2 \times \mathbb{S}^2)$. Remember that g is a function on the set of feature pairs and Δ embeds the epipolar constraint. In some sense we are computing the overlap or intersection between point pairs in g with epipolar great circles in Δ. In fact, we are searching for the rotation pair which maximizes this overlap. The general problem of signal correlation has been approached successfully in other domains. The convolution properties of functions on various groups and homogeneous spaces have shown that it is often easier to compute the spectral components of a correlation function like $G(R_c, R_t)$ than it is to generate the function samples directly in the spatial domain. To get a clearer understanding of how we can compute our integral in such a fashion, we can explore the simpler problem of maximizing the correlation between two functions defined on the unit sphere $\mathbb{S}^2$. In this setting we will compute

$$G(R) \;=\; \int f(p)\Lambda_R h(p)dp, \;\; f, h \in \mathcal{L}^2(\mathbb{S}^2) \quad (6)$$

by generating the spectral coefficients of $G(R)$. This approach naturally gives rise to three questions: *(1) How can we compute the Fourier transform of $f \in \mathcal{L}^2(\mathbb{S}^2)$? (2) How does the spectrum of f change under a rotation $\Lambda_R f$? (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. Readers are referred to [5] for a comprehensive exposition of the spherical Fourier transform.

As the solution to the Laplacian restricted to the circle generates a basis for periodic functions on the line, the spherical harmonic functions Y_m^l form an orthonormal basis for spherical functions. There exist $(2l+1)$ such harmonics for each degree l ($m = -l \ldots l$), and they are defined as

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

where $P_m^l(\cos\theta)$ are associated Legendre polynomials. This basis gives rise to the Spherical Fourier Transform (SFT):

$$\begin{aligned} f(p) &= \sum_{l\in\mathbb{N}} \sum_{|m|\le l} \hat{f}_m^l Y_m^l(p) \qquad (7) \\ \hat{f}_m^l &= \int_p f(p)\overline{Y_m^l}(p)dp \qquad (8)\end{aligned}$$

Two very important properties of the spherical harmonic functions are their orthogonality and their relationship under rotations:

$$\int_p Y^l_m(p)\overline{Y^n_k(p)}dp = \delta_{ln}\delta_{mk} \qquad (9)$$

$$\Lambda_R Y^l_m(p) = \sum_{|k|\le l} Y^l_k(p)U^l_{km}(R) \qquad (10)$$

The U^l are the unitary matrix representations of the transformation group $SO(3)$. This last relationship (10) is important because it helps answer our second question. We will write $\hat{f}^l, Y^l$ without the subscript m to denote the vector of $(2l+1)$ orders for a given degree l. With this notation we can express the inverse SFT (7) for functions undergoing a rotational shift:

$$\begin{aligned}\Lambda_R f(p) &= \sum_l (\Lambda_R Y^l(p))^T \hat{f}^l \\ &= \sum_l Y^l(p)^T U^l(R)\hat{f}^l \qquad (11)\end{aligned}$$

As the unitary matrices U^l are the group representations of $SO(3)$, they form a basis for a Fourier transform on the rotation group:

$$f(R) = \sum_l \sum_{|m,k|\le l} \hat{f}^l_{mk} U^l_{mk}(R) \qquad (12)$$

$$\hat{f}^l_{mk} = \int_R f(R)\overline{U^l_{mk}(R)}dR \qquad (13)$$

The matrix elements of U^l are given as

$$U^l_{mk}(R) = e^{-im\alpha}P^l_{mk}(\cos\beta)e^{-ik\gamma}, \qquad (14)$$

where P^l_{mk} are the generalized Legendre polynomials.

We now have the mechanisms in place to answer our third question. Replacing $f(p)$ and $\Lambda_R h(p)$ with their Fourier transforms we have

$$G(R) = \sum_l \sum_{|m,k|\le l} \overline{\hat{f}^l_m}\hat{h}^l_k U^l_{mk}(R) \qquad (15)$$

From the orthogonality property

$$\int_R U^{l_1}_{m_1k_1}(R)U^{l_2}_{m_2k_2}(R)dR = \delta_{l_1l_2}\delta_{m_1m_2}\delta_{k_1k_2}$$

the SO(3) Fourier transform of $G(R)$ is simply

$$\hat{G}^l = \overline{\hat{f}^l}(\hat{h}^l)^T \qquad (16)$$

In conjunction with the inverse $SO(3)$ Fourier transform (12), this last equation shows that we can obtain the samples of $G(R)$ directly from the pointwise multiplication of the Fourier coefficients of f and h.

INPUT

1. A pair of spherical images I_1, I_2

OFFLINE

1. Compute the Fourier transform $\hat{\Delta}$ of Δ from (18).

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 (18).
4. Generate the 5D coefficient space $\hat{G}^{l_1l_2}_{m_1m_2k_1-m_2}$ from $\hat{g}$ and $\hat{\Delta}$ as described in (19).
5. Using inverse Fourier transforms (12) 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. Relative orientation between cameras is $R = R_t R_c^T$.
8. Direction of translation is $T = R_t e_3$.

Figure 2. The full motion estimation algorithm.

As expected, this theory extends directly to functions on $\mathbb{S}^2 \times \mathbb{S}^2$, where the “rotation” comes from the product group $SO(3) \times SO(3)$. The Fourier transform for any function $f \in \mathcal{L}^2(\mathbb{S}^2 \times \mathbb{S}^2)$ is given as

$$f(p,q) = \sum_{l_1l_2m_1m_2} \hat{f}^{l_1l_2}_{m_1m_2} Y^{l_1}_{m_1}(p)Y^{l_2}_{m_2}(q) \qquad (17)$$

$$\hat{f}^{l_1l_2}_{m_1m_2} = \int_p\int_q f(p,q)\overline{Y^{l_1}_{m_1}(p)Y^{l_2}_{m_2}(q)}dpdq \qquad (18)$$

The spectrum of $G(R_c, R_t)$ from (5) can be obtained from the Fourier transforms of g, Δ:

$$\hat{G}^{l_1l_2}_{m_1m_2k_1k_2} = \overline{\hat{f}^{l_1l_2}_{m_1k_1}}\hat{\Delta}^{l_1l_2}_{m_2k_2} \qquad (19)$$

Up to this point we have treated camera motions with rotation pairs $(R_c, R_t) \in SO(3) \times SO(3)$. However, the direction of translation obtained from $t = R_t e_3$ is independent of the first applied rotation from R_t, so we fix $R_t = R(0, \beta, \gamma)$. In effect, the rotation R_t is explicitly a two-parameter rotation. This characteristic is reflected in our formulation since $\hat{G}$ is nonzero only if $m_2 = -k_2$. We are only interested in the coefficients $\hat{G}^{l_1l_2}_{m_1m_2,k_1,-m_2}$, which constitute the five-dimensional Fourier space of our camera motions. The resulting inverse Fourier transform required to obtain the samples of G is also only five dimensional. A summary of the full ego-motion estimation algorithm is presented in figure (2).

5 Experiments

In this section we will present the results of the motion estimation algorithm on real image sequences as well as a simulated result for detecting multiple motions in the scene. Before presenting the results we will address some practical considerations regarding spherical image acquisition and discrete Fourier transforms.

5.1 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. The projection model of a central catadioptric system is equivalent to a spherical projection followed by a projection onto the plane [7]. 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$^{\text{TM}}$[12]. The mirror's field-of-view is 212° so the camera captures slightly more than a hemisphere of information. The images from this system are mapped to a uniformly sampled polar grid. Figure (3) shows a sample catadioptric image obtained from a parabolic mirror and its corresponding projection onto the sphere.

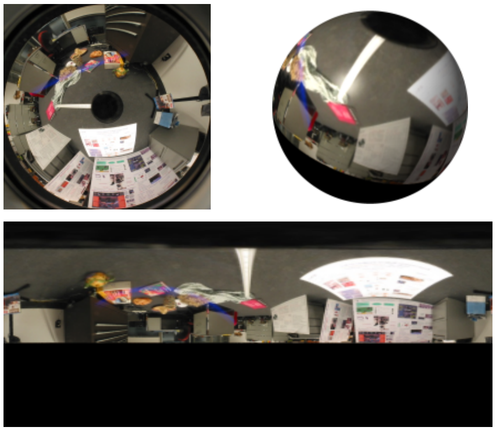

Figure 3. Top Left: a parabolic catadioptric image. Bottom: the corresponding spherical image on a uniformly sampled polar grid. Top Right: the spherical image as it would appear on the surface of the sphere

5.2 Discrete Fourier transforms

Until now we have only discussed the spherical and $SO(3)$ Fourier transforms in regards to continuous functions. However, our spherical images and Radon space $G(R_c, R_t)$ are discrete functions. In order to compute the SFT of a spherical image residing on a uniformly sampled polar grid, we can use a fast $O(L^2 log^2 L)$ algorithm developed by Driscoll and Healy [5], where L is the bandlimit of the signal being transformed. A similar separation-of-variables approach exists for a fast $SO(3)$ Fourier transform in $O(L^3 log^2 L)$ [10].

5.3 Results

We proceed to show experimental results of our algorithm tested on a sequence of real omnidirectional images. 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 immediate issues which would prevent a naive implementation from being operative. First is the volume of outliers. As the outliers in the set of feature pairs between two typical images is over 99%, 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. 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 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 figure (4). Notice in the figure that the translational slice shown depicts a peak at $R_t(0, \frac{\pi}{2}, \pi)e_3 = -X$. Although it is clear that translation along both $\pm X$ will satisfy the epipolar constraint, what may be surprising is that there is not also a peak at $+X$. This happens because R_c is a composite rotation of both rotational and translational terms, and so (R_c, R_t) and $(R_c, -R_t)$ do not represent the same motion.

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 figure (5). 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 figure (6). 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 the Radon's motion estimation.

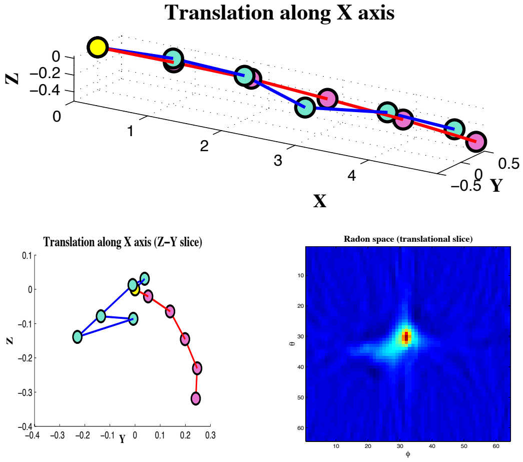

Figure 4. Top: the estimated trajectory of the camera. In blue (light) is the Radon estimation, in red (dark) is the RANSAC computation, and the yellow circle marks the starting position. Bottom Left: An Z-Y slice showing the deviation of the estimated positions from the X axis. Bottom Right: the R_t slice of the grid G where the maxima was found.

5.4 Multiple motions

One aspect of our algorithm we have only briefly touched upon is the significance of treating feature pairs independently. This is critical because while outlying feature pairs may contribute to incorrect solutions, they cannot detract from or perturb the value of the integral at the position of the correct solution. The effect, besides making our algorithm robust to outliers, is that if there are multiple moving objects in a scene, the feature matches from the individual objects will contribute to their respective motions. Thus, our algorithm, without having to be altered, can detect multiple motions of moving objects in a scene.

We simulate two moving objects in an otherwise static scene. Figure (7) shows a caricature of the types of scenes we considered for this simulation. The two objects each have 150 features. These features project onto the spherical image as gaussian blobs. A simple sum-of-squared differencing is used to generate the similarity function g. We incrementally deform the features by randomly replacing the σ with a new one from the existing pool (this simulates introducing erroneous feature matches while simultaneously reducing correct matches). Figure (7) shows on the bottom one of the two translational slices of the motion space.

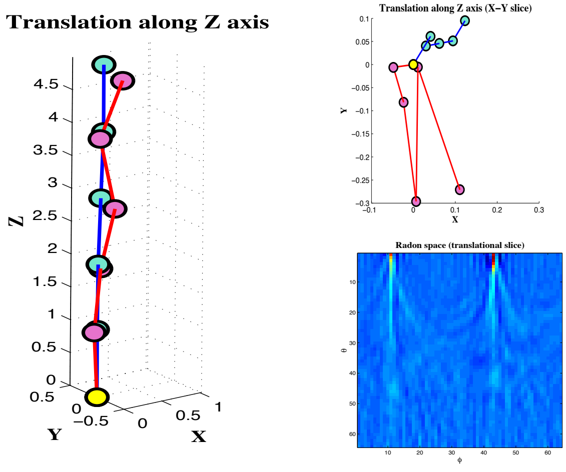

Figure 5. Left: the estimated trajectory of the camera. In blue (light) is the Radon estimation, in red (dark) is the RANSAC computation, and the yellow circle marks the starting position. Top Right: An X-Y slice showing the deviation of the estimated positions from the Z axis. Bottom Right: the R_t 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).

Although both motions are correctly estimated when 20% of the features are deformed, the peaks are clearly disintegrated by the time 30% of the features have been affected.

6 Conclusion

We have presented a novel approach for the computation of 3D-motion from two views without correspondences. It is based on a 5D-search in the motion parameter space. Given today's computing power it is not the search 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 function can be written as a Radon-transform which is known to become a convolution 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-less approaches. Naturally, our approach can handle both outliers and multiple rigid motions.

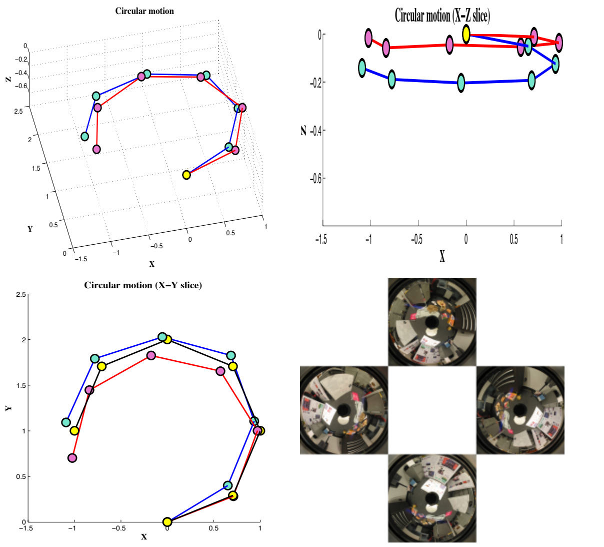

Figure 6. A camera moving along a circular path. Top Left: In blue (light) is the Radon estimation, in red (dark) is the RANSAC. Top Right: The X-Z slice showing the deviation from the plane of the turntable. Bottom Left: An overhead view. The yellow circles are the observed ground truth positions of 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.

It can be easily cast in a maximum likelihood framework. Our approach can be modified to incorporate a normalization of the epipolar constraint that removes the bias in translation direction.

References

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

[2] K. Daniilidis and M. Spetsakis. Understanding noise sensitivity in structure from motion. In Y. Aloimonos, editor, *Visual Navigation*, pages 61–88. Lawrence Erlbaum Associates, Hillsdale, NJ, 1996.

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

[4] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. Structure from motion without correspondence. In *CVPR*, Hilton Head Island, SC, June 13-15, 2000.

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

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

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

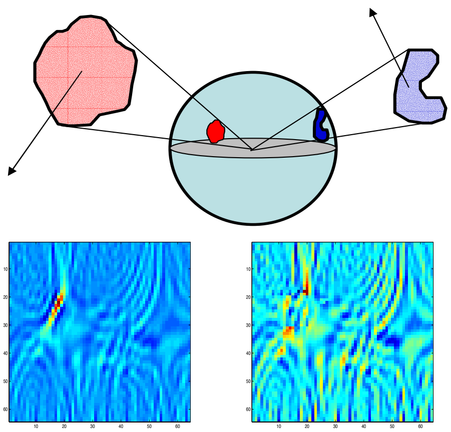

Figure 7. Top: A caricature describing the simulations for multiple motion detection. Our algorithm was tested with two objects moving independently in a static scene. Bottom Left: One of the two translational slices with 20% deformed features. Bottom Right: A translational slice with 30% deformed features.

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

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

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

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

[12] S. Nayar. Catadioptric omnidirectional camera. In *IEEE Conf. Computer Vision and Pattern Recognition*, pages 482–488, Puerto Rico, June 17-19, 1997.

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

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

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

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

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

[18] R. Vidal and Y. Ma. A unified algebraic approach to 2-d and 3-d motion segmentation. pages 1–15, 2004.

[19] Y. Wexler, A. Fitzgibbon, and A. Zisserman. Learning epipolar geometry from image sequences. In *IEEE Conf. Computer Vision and Pattern Recognition*, Wisconsin, June 16-22, 2003.