Integration of Visual and Inertial Information for Egomotion: a Stochastic Approach

Justin Domke and Yiannis Aloimonos
Computer Vision Laboratory, Dept. of Computer Science
University of Maryland, College Park, MD 20742 USA
{*domke,yiannis*}*@cs.umd.edu*

Abstract*— We present a probabilistic framework for visual correspondence, inertial measurements and Egomotion. First, we describe a simple method based on Gabor filters to produce** *correspondence probability distributions. Next, we generate a noise model for inertial measurements. Probability distributions over the** *motions* **are then computed directly from the correspondence distributions and the inertial measurements. We investigate combining the inertial and visual information for a single distribution over the motions. We find that with smaller amounts of correspondence information, fusion of the visual data with the inertial sensor results in much better Egomotion estimation. This is essentially because inertial measurements decrease the "translation-rotation" ambiguity. However, when more correspondence information is used, this ambiguity is reduced to such a degree that the inertial measurements provide negligible improvement in accuracy. This suggests that inertial and visual information are more closely integrated in a compositional sense.**

I. Introduction

In this paper, we address the problem of Egomotion estimation from visual and inertial measurements, a basic problem and a prerequisite for any navigational competence. Given two images, and two inertial measurements taken simultaneously, we wish to estimate the Egomotion of the sensor rig between the two frames. The standard way to use visual information for this task is to establish correspondences between the two images. As is well known, a few correspondences known with perfect precision suffice to find the exact Egomotion. In practice one cannot do this, for two reasons. First, the problem of establishing correspondences between two images is in general unsolvable- in the best case, there may be a few exceptional points where this may be done. Secondly, correspondences- even if they could somehow be manually checked for correctness- are known only with a finite precision. If this finite precision is taken into account, a small number correspondences yield a rather large group of possible motions- essentially because it is difficult to disambiguate image motion due to rotation and translation.

For inertial measurements, if the gravity vector were exactly known in each frame, the degrees of freedom for the rotation would be reduced from 3 to 1. If this reduction were perfect, the translation-rotation ambiguity would be markedly reduced. However, since inertial measurements are always corrupted by noise, the true rotation will not be exactly compatible with this constraint. Still it is natural to observe that if the inertial sensor can give independent information about the rotation, this can be used to combat the rotation-translation ambiguity in vision-based Egomotion estimation.

We propose to treat both correspondence and inertial measurements to be, in general, unmeasureable. Instead, given the visual and inertial data, we establish *probability distributions* over the correspondences and gravity vectors. We are then able to directly calculate the probability of different motions, with out committing to specific values for these underlying, uncertain quantities. We will see that this strategy makes it possible to extract a great deal of correspondence information from the images- much more than approaches which limit themselves to feature points. We then present experiments that arrive at a somewhat counterintuitive result- though if relatively small amounts of correspondence information are used, inertial measurements allow for more accurate Egomotion, it is possible to extract so much correspondence information from images that inertial measurements provide virtually no increase in performance. We conclude with a general discussion of what role inertial sensors might play in robotics.

There are several novel aspects of this work. First, our method of computing correspondence probability distributions from the phase of tuned Gabor filters is new. A second contribution is our method of calibrating the inertial sensor for its noise profile, and then using this to interpret inertial measurements probabilistically. Third is our method to compute probabilities of different motions directly from probability distributions of correspondence. We point out that our method is essentially a probabilistic phrasing of least-squares epipolar minimization, generalized to the case of *probability distributions* of correspondence. Fourth, we present the combined visual/inertial probabilistic framework, and experiments suggesting that if visual information is used optimally, inertial sensors may not much help to improve Egomotion estimation.

A. Related Work

As Egomotion estimation is one of the oldest and most widely researched areas of computer vision, the reader is referred to a survey [9] or a recent textbook [12] for a summary. The first category of more closely related papers discuss correspondenceless visual Egomotion techniques. Several algorithms have been proposed based on the computation of the normal flow [10]. Though these algorithms will not suffer from the aperture problem, they do not address problems

such as repetitive structure, and are therefore tangential to this work. Wexler et al. [13] present a method which aggregates information over multiple image pairs to learn the epipolar geometry. Dellaert et al. [2] present an algorithm which iteratively computes probabilities over both correspondence and motion through the Expectation-Maximization framework. The principal drawback of this algorithm is the possibility of getting 'stuck' in a poor solution. Unrelated to Egomotion, Rosenberg and Werman [14] use probability distributions of correspondence for object tracking.

The second category of relevant literature is regarding the integration of visual and inertial sensors. There is the Inervis workshop [8] dedicated to this. Lobo and Diaz [1] explore the use of inertial data with visual sensors, including thorough references to earlier work. Spanning both categories, Makadia and Daniilidis present a technique for panoramic imaging devices where inertial measurements are used to reduce the unknown inertial parameters from 3 to 1, followed by a correspondenceless Hough-transform search for the best parameters in the remaining 3 dimensional motion space [7].

In our discussion of the accuracy of Egomotion estimates, we will focus on the 'translation-rotation ambiguity'- the fact that when estimating motion from a finite field of view camera, translation and rotation are easily confounded. This has been noticed repeatedly in practice, and addressed in theoretical analyses [5] [4].

Supporting our approach's philosophy, Thrun [3] argued broadly for probabilistic perception in Robotics, since any system must deal explicitly with the uncertainty in its measurements to perform optimally.

II. Correspondence

It has long been known that in general, correspondences cannot be reliably established between two arbitrary images. The aperture effect is the most widely discussed cause of this. Simply stated, if correspondence is sought for a point lying along a straight edge, information is only available to constrain the point to lie along the corresponding edge in the other image. This constraint is known as the normal flow, and Egomotion algorithms exist that estimate motion directly from it [10]. Also common in practice is the problem of repetitive texture. Here, correspondence can be constrained to a group of possibly disjoint points- lack of texture may be thought of as an extreme case of this. These problems affect different parts of the images to different degrees. Feature detectors may be thought of as locating points that are relatively immune to them.

Since weaker forms of correspondence such as normal flow will give up information unnecessarily at points that do not happen to show any ambiguity, they do make use of all available image information. Similarly, restricting consideration to feature points severely limits the number of possible correspondences. As we will discuss later, large amounts of correspondence information are vital for accurate Egomotion estimation. We propose that by instead computing *correspondence probability distributions*, all of these problems are overcome. If a point is subject to some particular ambiguity, the distribution can represent it. At the same time, if the correspondence for a point is clear, the distribution need not give up information unnecessarily. Distributions may be computed for any point, representing all of the information about correspondence to be found there.

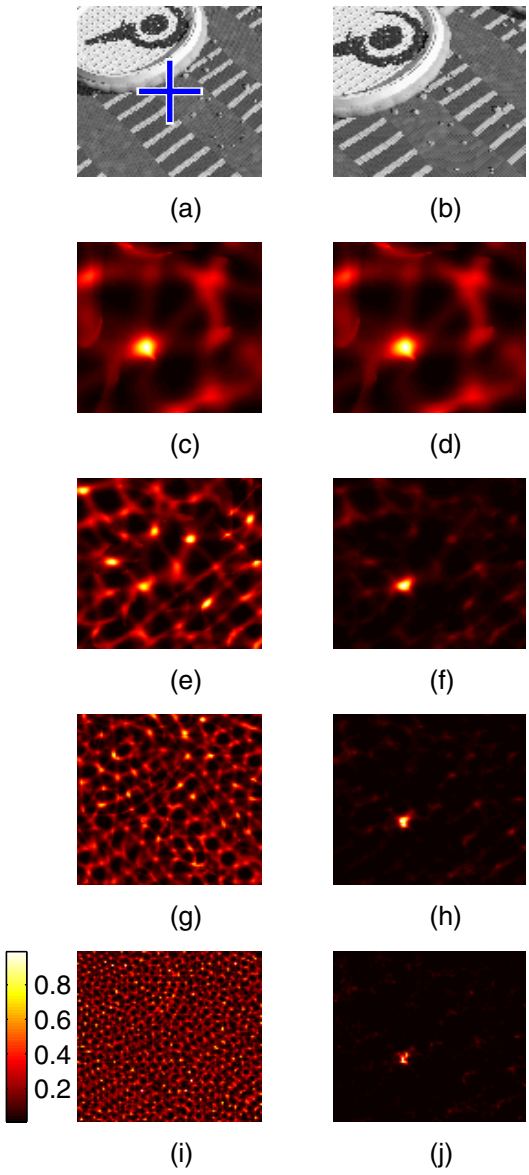

Fig. 1. The computation of a correspondence distribution. (a) first image. (b) second image, over which correspondence is being considered. (c)-(i) distributions for specific, decreasing scales, each with all orientations. (d)-(j) distribution considering all previous scales. (j) final distribution.

Our method of computing correspondence distributions is based on Gabor filters. Namely, we exploit the fact that for a filter with a given orientation and scale, matching points will have matching phase. We develop a probability distribution for each orientation and scale, and combine these for the final distribution. Suppose we are attempting to estimate the probability that a pixel s matches most closely to a pixel $\hat{q}$.

For the filter with orientation γ and scale l, denote the phase by $\phi_{l,\gamma}$. We take the probability that s and $\hat{q}$ match to be proportional to $\exp(-|\phi_{l,\gamma}(s)-\phi_{l,\gamma}(\hat{q})|^2)+1$. Combining the information over all filters, then,

$$\rho_s(\hat{q}) \propto \prod_{l,\gamma}(\exp(-|\phi_{l,\gamma}(s) - \phi_{l,\gamma}(\hat{q})|^2) + 1).$$

Thus, points whose phase is very close will have much higher probability. Since we are computing a probability over a discrete set of points, it cannot be insisted that the phase match exactly. Furthermore, to increase robustness to noise, we add the constant of 1, limiting the influence of any single filter. The computation of an example distribution is illustrated in Fig. 1. Parts (c)-(j) show the probability that the point marked in (a) corresponds to each possible location in (b). Probabilities are encoded as color. It can be seen that the large scale Gabor filters (i.e. (c)) provide different distributions that the small scale filters (i.e. (i)). Nevertheless, it is shown in the right column that the combination of all filters leads to an excellent distribution. Now, since s will not generally correspond exactly to a point with pixel coordinates, we use the following expression as a kind of 'interpolation' to reflect the probability that s corresponds to an *arbitrary* point q.

$$\rho_s(q) \propto \max_{\hat{q}} \rho_s(\hat{q}) \exp(-||\hat{q} - q||^2) + \alpha$$

As we will discuss later for the case of known correspondence, this Gaussian distribution is implicitly assumed when minimizing over the least-squared epipolar distance. Thus this expression may be thought of as a natural generalization to the case of probability distributions over correspondence.

The constant of α is added to reflect the possibility that the correspondences computed are not accurate. This would be the case, for example, if the point corresponding to s were occluded in the second image. Adding this constant is equivalent to taking a certain probability that the image information is unreliable, in which case a 'flat' distribution is appropriate.

In our implementation, we found it convenient to use a low threshold $\rho_{\min}$ where if $\rho_s(\hat{q}) < \rho_{\min}$ it is set to zero and neglected from further consideration. When $\rho_{\min} < \alpha$ this approximation has negligible impact on results, while resulting in better performance.

III. Inertial Measurements

In this paper, we will consider rotations parameterized by a length 3 vector ω. We will speak of a rotation matrix $R(\omega)$. By this we mean the matrix which rotates all points by an angle $|\omega|$ about the unit vector $\omega/|\omega|$. Now, suppose the gravity vector is measured perfectly in two frames, with measurements g_1, and g_2. Since gravity is unchanging, this presents a constraint on the motion, $g_2 = R(\omega)g_1$. Note that this constrains only two of the three rotational degrees of freedom, since if $g_2 = R(\omega)g_1$, then $g_2 = R(cg_2)g_2 = R(cg_2)R(\omega)g_1$ for any scalar c. If the vectors g_1 and g_2 are normalized to have length one, we may write this constraint as $(R(\omega)g_1) \cdot g_2 = 1$.

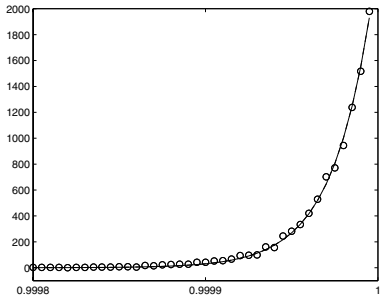

Fig. 2. Inertial Error Model

For any real sensor, however, the measurements will inevitably be corrupted by noise. If the equation above is used as a 'hard' constraint, it may very well result in a worse Egomotion estimate. To make robust and optimal use of the inertial data, we must have a realistic noise model. To attain an estimate for the error in our inertial sensor, we recorded 10,000 consecutive measurements while the sensor was stationary. Measurements were normalized so that all measurements had length 1. We then calculated $g_1 \cdot g_2$ for each consecutive measurement pair. These measurements are plotted in a histogram with bin size .000005 along with a fitting line in Fig. 2. There, circles indicate the histogram values; the solid line indicates the fit. The fitting line is $p(g_1 \cdot g_2) = 2402(g_1 \cdot g_2)^{4.374e4}$. This suggests that given a measurement g_1 for gravity in one frame, and g_2 in another, we can estimate the probability that ω is the rotation between the frames by $\rho(\omega) \propto (R(\omega)g_1 \cdot g_2)^{\mu}$, where $\mu = 4.374e4$. We should note that this is a noise model for our specific inertial sensor, a 3DM-GX1, manufactured by MicroStrain, Inc. Since different sensors are sure to have different noise properties, individual calibration is unavoidable.

IV. Sensor Rig Calibration

As pictured in Fig. 3 on our sensor rig consists of a camera, rigidly attached to an inertial sensor. Though an approximate answer could be found from physical measurements, we take the coordinate systems defined by the two sensors to be subject to an arbitrary rigid transformation. Since the direction of gravity will be unaffected by the translation, we can neglect it completely, and focus only on finding a rotational

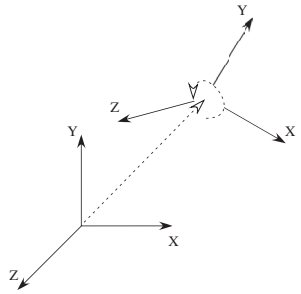

Fig. 3. The coordinate systems for the camera and inertial sensor differ by an arbitrary rigid transformation.

matrix $C(\omega_c)$ which can transform gravity vectors in the inertial frame to the camera frame. With this done, we can use the inertial measurements to find the Egomotion in the camera's frame. We captured an outdoor sequence consisting of approximately 50 frames, each frame consisting of both visual and inertial data. In this sequence we attempted to keep the translational motion of the camera to a minimum. As all objects in the scene were very distant relative to the translation, essentially all image motion was due to the camera rotation. We then found rotation matrices R_i for all frame pairs from the visual data. For simplicity, this may be thought of as being done manually. The calibration matrix C can then be found by searching for rotational parameters w_c such that $\forall i \in 1, ..., N-1, C(\omega_c)g_{i+1} \approx R_i C(\omega_c)g_i$. Specifically, we used a gradient descent method to search for ω_c minimizing:

$$\sum_{i=1}^{N-1} ||C(\omega_c)g_{i+1} - R_i C(\omega_c)g_i||^2$$

V. Egomotion

A. Egomotion from Vision

Given a correspondence probability distribution for a *single* point s, we take the probability of a given motion to be proportional to the maximum probability correspondence which exactly satisfies the epipolar constraint:

$$\rho_s^V(t,\omega) \propto \max_{q:qEs=0} \rho_s(q)$$

Here, $E = [t]_\times R(\omega)$ is the Essential Matrix corresponding to the motion (t,ω) [12]. One can substitute our earlier formula for $\rho_s(q)$ to obtain:

$$\rho_s^V(t,\omega) \propto \max_{q:qEs=0} \max_{\hat{q}} \exp(-||q-\hat{q}||^2) + \alpha$$

Now, observe that this is in fact equal to:

$$\rho_s^V(t,\omega) \propto \max_{\hat{q}} \exp(-||\hat{q}^T l_{(E,s)}||^2) + \alpha$$

Where $l_{(E,s)}$ is the line Es normalized so that for any point r, the distance from the line to the point s on the image plane is simply $r^T l_{(E,s)}$:

$$l_{(E,s)} = \frac{Es}{\sqrt{(E_1 s)^2 + (E_2 s)^2}}$$

Combining the information over all points into $\rho^V(t,\omega) \propto \prod_s \rho_s^V(t,\omega)$ then yields the final probability in form in which it is calculated:

$$\rho^V(t,\omega) \propto \prod_s (\max_{\hat{q}} \rho_s(\hat{q}) \exp(-(\hat{q}^T l_{(E,s)})^2) + \alpha)$$

It may be argued that the approach we have outlined here is heuristic. For example, why is the Gaussian distribution used for points with out pixel coordinates? Indeed, there are numerous ways to develop probability distributions of motion from images. However, our approach may be considered as a reasonable generalization of previous work on the grounds that, with known correspondences, maximizing our function is exactly equivalent to minimizing the least-squared epipolar distance, as we will now show. Suppose that we have known matches- each point s_i is known to correspond to the pixel $\hat{q}_i$. In this case, we would have $\rho_{s_i}(\cdot) = 1$ for $\hat{q}_i$ and 0 for all other points. Furthermore, since the matches are known to be correct, α should be set to 0. Therefore,

$$\arg\max_{t,\omega} \rho^V(t,\omega) = \arg\max_{t,\omega} \prod_i (\rho_{s_i}(\hat{q}_i) \exp(-(\hat{q}_i^T l_{(E,s)})^2))$$

$$= \arg\max_{t,\omega} \prod_i \exp(-(\hat{q}_i^T l_{(E,s)})^2))$$

Since the motion which maximizes the right side will also maximize its logarithm, we obtain the exact expression for the least-squares epipolar distance:

$$= \arg\min_{t,\omega} \sum_i (\hat{q}_i^T l_{(E,s)})^2$$

B. Egomotion from Inertial Sensors

The probability distribution over the motions is given directly by the noise model described in section III:

$$\rho^I(t,\omega) \propto (R(\omega)g_1 \cdot g_2)^\mu$$

C. Sensor Fusion

The final probability for a given motion is given by the assumption that the information given by vision and inertial sensors is independent. We can simply combine the probability distributions to obtain our full probability distribution over the space of motions:

$$\rho(t,\omega) \propto \rho^V(t,\omega)\rho^I(t,\omega)$$

D. Optimization

Given a single motion, (t,ω), its probability can be very quickly computed. Still, because there are 5 degrees of freedom, computing a full motion probability distribution is problematic- computational considerations demand such a coarse sampling of each dimension that the entire 'peak' of the distribution may be missed. In our experiments, we will maximize the motion function through a simple heuristic optimization. First, random sampling (t on the sphere with $|t| = 1$, ω such that $|\omega| \leq .1$) is used at approximately 2500 points. Next, the Nelder-Mead simplex search method is used at the 25 highest scoring samples. The final maximum probability sample found is taken as the result. In practice, we found that several of the 25 searches resulted in very close answers, suggesting that missing the global maximum altogether is unlikely. This is consistent with results reported elsewhere for Egomotion techniques using nonlinear functions [4] [11] suggesting there will be several (but only several) local minima. Although this is in a sense a 'brute-force' maximization, in practice the slowest part of our technique is the computation of the correspondence distributions. In our implementation, the function is maximized in a few seconds when using a small amount of correspondence information, and in the order of a

minute even when using a very large amount of information. During the search, t is parameterized by the azimuthal and polar angles on the sphere. It is possible that a practical real-time system could be constructed to maximize the function in real time, by computing the correspondence and motion distributions in parallel using specialized hardware.

VI. Experiments

A. Synthetic Measurements

As a first experiment, we generated two synthetic images with known egomotion. Next, we generated 10,000 gravity vectors, under the assumption that that gravity was along the z axis. These gravity vectors were polluted with noise so that they produced the same distribution as shown in Fig. 2. We admit that synthetic images and inertial measurements are somewhat unsatisfying, but this is the most practical way to generate a sequence with the exact known motion. Then, to provide a comparison for our approach, we manually extracted 50 pixel-accurate correspondences between the two images. Since there are a large number of methods which attempt to automatically establish feature correspondences, we select them by hand to present an upper-bound on their performance. In Fig. 4 we compare the performance of 4 techniques: First, the algorithm run on the hand matches- equivalent to least-squares epipolar minimization. Second, we show the algorithm using both hand matches, and inertial measurements. Third is the algorithm using only the correspondence probability distributions, while fourth is the correspondence distributions with inertial information. For each point shown, means were taken over 100 trials. Translational error is computed as $\sum_{i=1}^{100} ||t_i - t_0||/100$, and rotational error as $\sum_{i=1}^{100} ||\omega_i - \omega_0||/100$, where (t_0, ω_0) is the ground truth motion. Each trial took a random subset of the appropriate correspondences, and random inertial measurements. In Fig. 5 and Fig. 6 the resulting solutions are shown 'projected' down into the two dimensions, t_x, and ω_y. A somewhat counterintuitive result is seen here. With small amounts of correspondence information, the inertial measurements greatly increase the accuracy of Egomotion estimation. However, when the very large amounts of correspondence information made available by computing probability distributions are used, the inertial sensor provides less and less a boost to performance, finally resulting in no apparent increase at all.

To aid in understanding what is happening here, in Fig. 7 we show probability distributions obtained with different amounts of correspondence information. With small amounts of correspondence info, a clear 'ridge' is seen, where a change of rotation is compensated with a change of rotation to yield a motion similarly consistent with the epipolar constraint. In this case, observe that the inertial information both reduces the 'ridge' and moves the peak closer to the correct answer. However if huge amounts of correspondence information are used, the 'ridge' is so small, that it entirely lies within the range of uncertainty for the inertial measurements. Here, the inertial sensor does almost nothing to increase the accuracy.

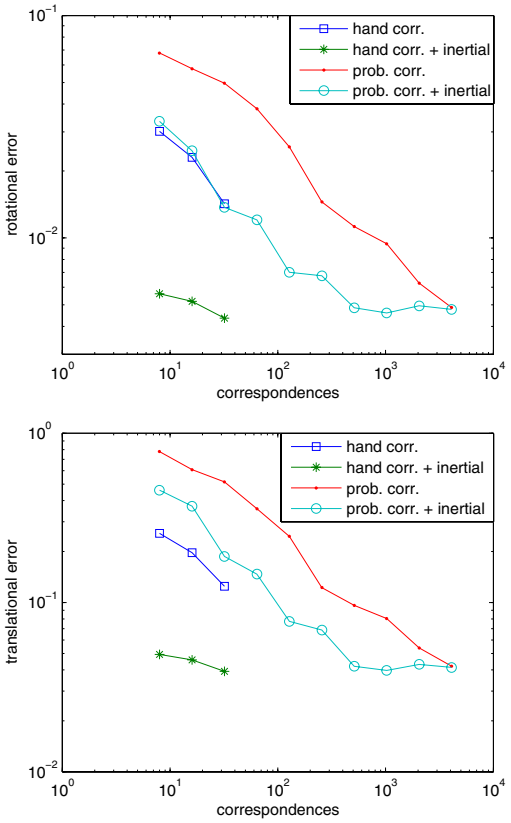

Fig. 4. Total errors for different numbers of correspondences or correspondence distributions.

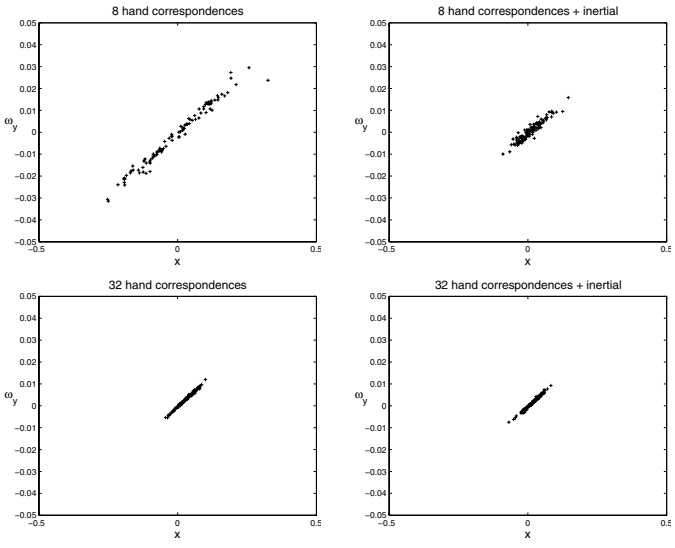

Fig. 5. Projected solutions for synthetic images, using hand matches.

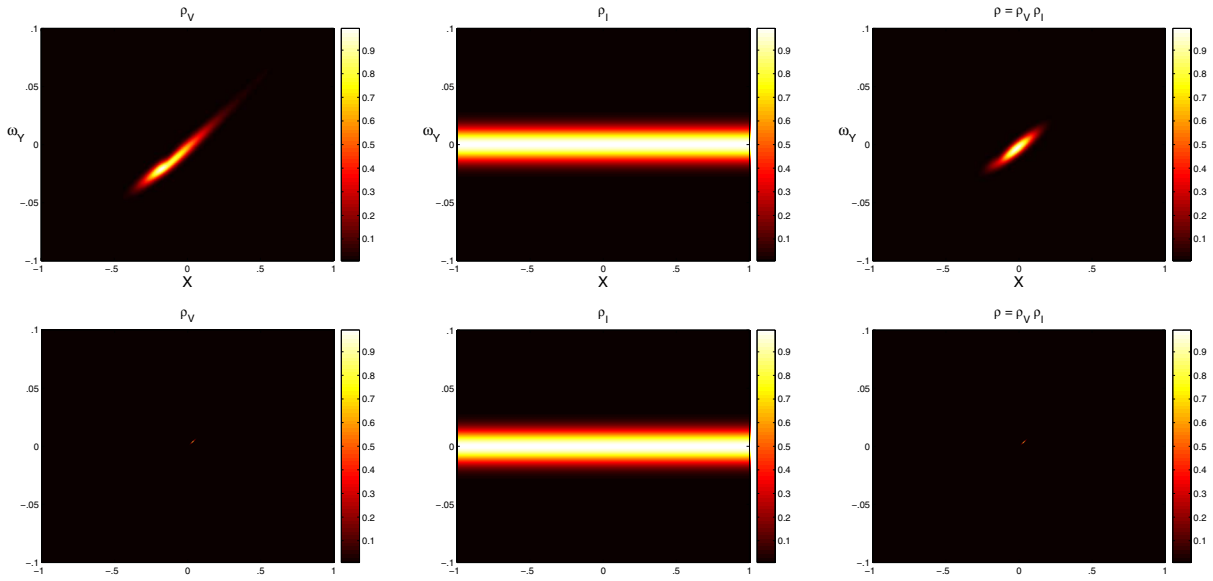

Fig. 7. 'Slices' of the motion probability distribution. Top Row: 10 hand-matches. Bottom Row: 5000 correspondence distributions. Left Column: distribution from vision alone. Center Column: distribution from inertial information alone. Right Column: combined distribution.

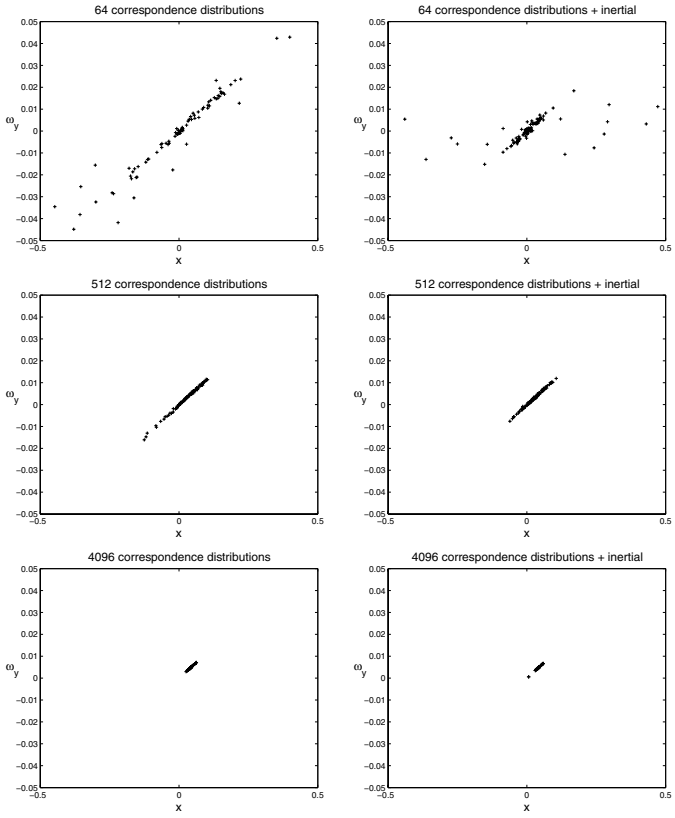

Fig. 6. Projected solutions for synthetic images, using the probabilistic algorithm.

B. Real Measurements

Results with real measurements from our sensor rig are shown in Fig. 8. Here, ground truth is not available, but as can be seen in Fig. 9 the projections found from different subsets of correspondences converge to a small area as the number of correspondences is increased. Again, we can see that when 4096 correspondence distributions are used, the inertial measurements only very slightly change the resulting Egomotion estimates.

VII. Discussion

In this paper, we have presented a probabilistic framework for the interpretation of visual and inertial measurements in the estimation of Egomotion. We have shown that by computing probability distributions of correspondence, very large amounts of correspondence information can be extracted from the images, leading to much more accurate Egomotion estimation. Though inertial sensors dramatically increase the accuracy of Egomotion estimates for small amounts of correspondence information, the inertial sensors provide virtually no benefit for the largest amounts. What are we to make of these results? We do not believe that they should be taken to suggest that inertial sensors are useless for Egomotion. On the contrary, given the ubiquity of inertial sensors in biological systems, we view this as merely suggesting more specific uses.

First, inertial sensors could provide computational advantages. Our technique essentially uses the visual and inertial sensors to independently estimate Egomotion estimation, combining information afterwards. Instead, they could be more tightly coupled- the inertial measurements guide the visual Egomotion process. If the framework described here were to be implemented in a practical real-time system, the portions

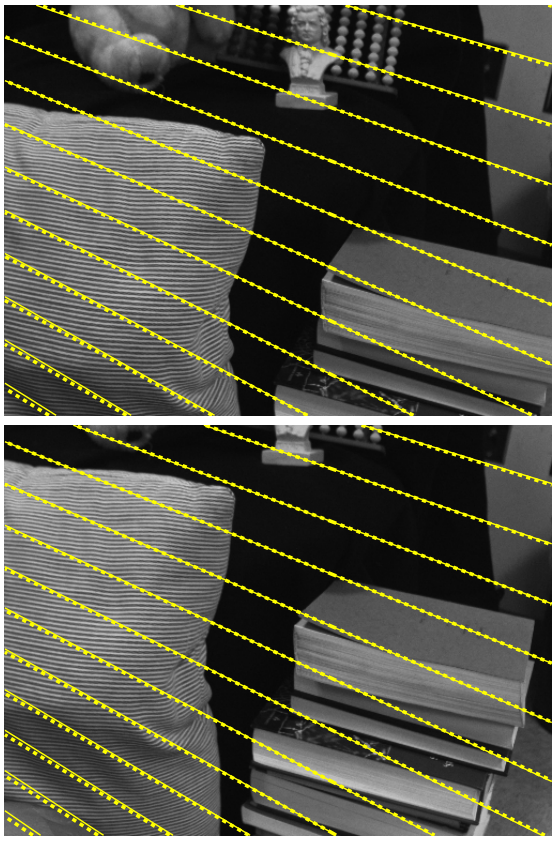

Fig. 8. Results for the probabilistic algorithm with two frames from the real sequence. Solid lines: Epipolar lines for visual information only. Dotted lines: Epipolar lines for Visual and Inertial information.

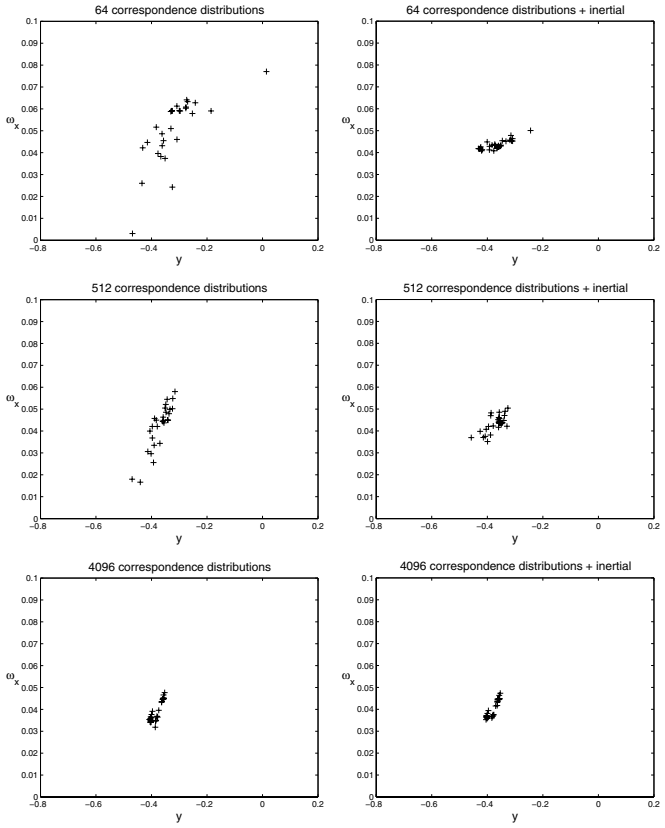

Fig. 9. Projected solutions for real images, using the probabilistic algorithm.

of the motion parameter space can be greatly 'pruned' by the inertial sensors, even if this does not result in a final change in accuracy. For example, if hardware were built to sample the motion space in parallel, the inertial measurements could be used to parameterize the portions of the space that were considered. Thus, the inertial sensors could enable use of a relatively small number of samples to represent the entire relevant motion space.

Second, inertial measurements could provide less direct information about motion. If there is independent motion in the scene, the inertial measurements could reject as incompatible any motions inconsistent with the gravity constraint. In this way, visual Egomotion would be possible even when most of the scene is taken up by independent motion. Furthermore, we should note that the way inertial measurements were used in this paper may not always be realistic. Inertial measurements do not capture gravity alone, but gravity *along with the acceleration of the sensor*. We might therefore use inertial sensors in the opposite of the way proposed here. Rather than using inertial sensors to find rotational information, the rotation could be found from vision, allowing the inertial sensors to compensate for gravity and find the *translational acceleration*. Future work should further investigate this tighter coupling of the use of different measurements.

REFERENCES

[1] Jorge Lobo, Jorge Dias. "Vision and Inertial Sensor Cooperation Using Gravity as a Vertical Reference," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 12, pp. 1597-1608, December 2003.

[2] F. Dellaert, S. Seitz, C. Thorpe, and S. Thrun. Structure from Motion without Correspondence,' *Proc. CVPR*, June, 2000.

[3] S. Thrun, Probabilistic algorithms in robotics, *AI Mag*., vol. 21, no. 4, pp. 93-109, 2000.

[4] J. Oliensis, The Error Surface for Structure from Motion, NEC TR, 2001.

[5] C. Fermüller and Y. Aloimonos. Observability of 3D motion, *International Journal of Computer Vision*, 37:43-63, 2000.

[6] G. Qian, R. Chellappa. Structure from motion using sequential Monte-Carlo methods, In *P*roc. ICCV, 2001.

[7] A. Makadia and K. Daniilidis. Correspondenceless ego-motion estimation using an imu. In *Proc. ICRA*, 2005.

[8] Workshop on Integration of Visual and Inertial Sensors, INERVIS, ICRA 2005, Barcelona Spain.

[9] J. Oliensis, A Critique of Structure from Motion Algorithms, *Computer Vision and Image Understanding*, 80:172- 214, 2000.

[10] T. Brodský, C. Fermüller, and Y. Aloimonos. Structure from motion: beyond the epipolar constraint, International Journal of Computer Vision, 37:231-258, 2000.

[11] T. Tian, C. Tomasi, and D. Heeger. Comparison of approaches to egomotion computation. In *Proc. CVPR*, 1996.

[12] Hartley, R. and Zisserman, A. 2004. *Multiple view geometry in computer vision*, Cambridge University Press: Cambridge, UK.

[13] A.Wexler, Y. Fitzgibbon and A. Zisserman. Learning epipolar geometry from image sequences. In *Proc. CVPR*, volume 2, pp. 209-216, 2003.

[14] Y. Rosenberg and M. Werman. Representing local motion as a probability distribution matrix applied to object tracking, In *Proc. CVPR*, pp. 654-659, 1997.