A Probabilistic Framework for Correspondence and Egomotion

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

Abstract

This paper is an argument for two assertions: First, that by representing correspondence probabilistically, drastically more correspondence information can be extracted from images. Second, that by increasing the amount of correspondence information used, more accurate egomotion estimation is possible. We present a novel approach illustrating these principles.

We first present a framework for using Gabor filters to generate such correspondence probability distributions. Essentially, different filters 'vote' on the correct correspondence in a way giving their relative likelihoods. Next, we use the epipolar constraint to generate a probability distribution over the possible motions. As the amount of correspondence information is increased, the set of motions yielding significant probabilities is shown to 'shrink' to the correct motion.

1. Introduction

Perhaps the single most pervasive structure in computer vision is that of correspondence - two points in different images that are said to correspond to the same point in space. Given a set of correct correspondences, powerful techniques exist to do many things - find camera egomotion, 3d depth, motion segmentation, etc. Thus most algorithms proceed by first matching points, and then using these correspondences to solve the problem at hand. Yet, it is well known that, in general, low level measurements do not provide sufficient information to match. This is not the paradox it might first seem to be. There are essentially 3 conditions resulting in correspondence being difficult to establish- repetitive structure in the scene, aliasing, and the aperture effect [1]. Feature detectors such as corner detectors [4] or SIFT features [5] may be thought of as algorithms that locate points in the scene that are relatively immune to these effects.

In this paper, we propose that a different structure could be used, namely, a *probability distribution* over the possible correspondences. There are several reasons to use such an approach. First, at those points in the scene for which correspondences are most easily found (e.g. non-repetitive feature points), we should expect that probability distribution to be nearly zero except at the true correspondence, meaning that no information needs to be given up at these points. In our experiments, feature points generally do yield localized probability distributions, though often so do points that would not be detected as feature points. Second, it is possible to represent arbitrary ambiguities in the correspondence, be they the result of aperture, repetitive structure, lack of texture, etc. Third, and perhaps most importantly, a probability distribution can be reliably found for *every point* in the scene. Though a point with a "spread out" distribution may provide weaker information than one with a sharp "peak", it is advantageous to make use of as much of the available information as possible.

Using these correspondence probability distributions leads naturally to a measure of the probability of different 3d motions. This measure is robust to occluded points and independent motion. We use the epipolar constraint to give an expression which is easily and quickly calculated. We will show that when a small number of correspondence distributions are used, a significant set of motions generally yield significant probabilities. However, because our framework gives us a very large number of correspondence distributions, they can all be used to reduce this set, yielding a very accurate egomotion estimate.

We first give a simple contrast invariant technique for calculating a correspondence probability distribution. Next, we show how these distributions may be used to calculate egomotion for a calibrated camera. We will present experiments showing that this egomotion technique is comparable in accuracy to a epipolar minimization algorithm based upon many manually extracted pixel-accurate correspondences. We will also show that this algorithm performs well in dynamic scenes, where objects in view violate the common assumption in egomotion algorithms that only the camera is moving.

1.1 Related Work

It is well known that correspondences cannot be reliably estimated from low-level measurements [7]. Simoncelli *et al*. [11] assume image gradients are corrupted by a Gaussian noise model, resulting in a probability distribution over the optical flow. This distribution is then used to estimate a single optical flow vector as output. Clocksin [1] estimates optical flow distribution functions for each point, and then uses spatiotemporal support regions to estimate more accurate (non-probabilistic) flow vectors each point.

Our approach to computing correspondence probability distributions is based on the phase of tuned Gabor filters. Phase has been widely used in the computation of stereo disparity [2] [10] as well as in one of the best performing optical flow algorithms [3]. We use the efficient Gabor filter implementation of Nestares *et al.* [8].

Egomotion and Structure from Motion are among the most heavily researched areas of computer vision research, and rather than attempting to summarize all references, the reader is referred to a survey [9]. The approach most similar to the one here is by Makadia *et al.* [6]. There, the authors use traditional feature points, but rather committing to an explicit matching, they search for a motion such that each feature point has a compatible point in the other image satisfying the epipolar constraint. Their approach can be phrased probabilistically. The principal difference with the current work is that we extract correspondence information for all points in the image, with out use of a feature detector. This means both that additional correspondence information is available, and that it is not necessary for the same point to be reliably detected as a feature. This drastically increased amount of correspondence information results in major increases in accuracy and robustness.

2. Correspondence Probability Distributions

Given a point s in one image, we would like to represent the probability that it corresponds to a point q in the next image. We should represent the probability that s moves to an arbitrary q, not necessarily with integer coordinates. We cope with this by first approximating the probability that s matches to a pixel $\hat{q}$, having integer coordinates. The probability that s matches to an arbitrary point is then represented via a Gaussian function. That is, we take the probability that s corresponds to q to be

$$\rho_s(q) = \max_{\hat{q}} \rho_s(\hat{q}) \exp(-||\hat{q} - q||^2),$$

where the points s, q and $\hat{q}$ are on the image plane, and $||\cdot||$ denotes the Euclidean norm. It will be seen later that this unusual form of interpolation simplifies the method.

Now, we want to find the probability that some pixel s corresponds most closely to another pixel $\hat{q}$. There are many possible ways to do this, but we follow many others in basing our approach on Gabor filters. These widely used filters can be tuned to different frequencies and orientations to provide a local measurement of phase. Correspondence is then estimated by exploiting the fact that phase will be nearly the same for corresponding points. Fleet [2] shows how the different filters form a voting scheme for stereo disparity. Our approach is similar, but more than ensuring the highest "score" for the most likely correspondence, we would like the scores to reflect the appropriate probabilities. Suppose the phase for the filter with orientation l and frequency ω at a point a is $\phi_{l,\omega}(a)$. We would like to consider points with very nearly matching phase to be likely to correspond. Simultaneously, any single filter, because of noise, may be unreliable. We therefore take the probability given by a single filter (l, ω) that s and $\hat{q}$ match to be proportional to $\exp(-|\phi_{l,\omega}(s) - \phi_{l,\omega}(\hat{q})|^2) + \beta$. The added constant of β is equivalent to taking a certain probability that the filter's information is wrong, perhaps because of occlusion or noise. Combining the probabilities over all filters then gives us

$$\rho_s(\hat{q}) = C_s \prod_{l,\omega} [\exp(-|\phi_{l,\omega}(s) - \phi_{l,\omega}(\hat{q})|^2) + \beta]$$

Where C_s is chosen so that $\sum_{\hat{q}} \rho_s(\hat{q}) = 1$. In all experiments shown, we have used $\beta = 1$. Some example distributions are shown in Figure 1. Though we will not focus on this here, we should note that the above approach only uses the phase of the Gabor filter response, and is thus highly contrast invariant.

In all results shown here, we have used Gabor filters with four orientations, and four frequencies. For the sake of computational efficiency, a low threshold can be used, where if $\rho_s(\hat{q}) < \rho_{\min}$, it is taken as equal to zero, and therefore removed from consideration. It is important to note that the approach we have outlined here will give unpredictable behavior when a point which is visible in the first image becomes occluded. The egomotion approach below does not attempt any filtering to remove these points. Nevertheless it is robust to this behavior, as well as being robust to independently moving areas.

3. Egomotion Probability Distribution

Given the correspondence probability distributions for all points, we would like to calculate the relative probabilities of different 3d motions. First, given a line l in homogeneous coordinates, we will need the minimum distance on the image plane between that line and a point p. If the line is normalized by taking $l \leftarrow \frac{l}{\sqrt{l_1^2 + l_2^2}}$, and p is normalized

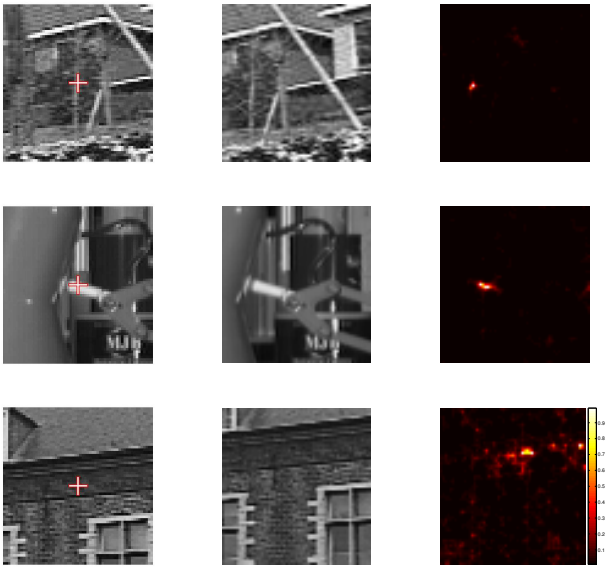

Figure 1. Example optical flow probability distributions. Left column: first image, with the point whose correspondence is being considered marked. Center column: second image. Right column: probability distribution over the points in the second image, with probability encoded as color.

by $p \leftarrow \frac{fp}{p_3}$, with f the focal distance, then the distance is simply $p^T l$.

Now, given the correspondence probability distribution for a single point s, we take the probability of a given motion hypothesis E to be the maximum probability $\rho_s(q)$ such that s and q satisfy the epipolar constraint, $qEs = 0$. To represent the fact that $\rho_s(q)$ may be wrong- if s becomes occluded, or belongs to an independently moving object- we add a constant α. This limits the influence or any single point to the egomotion probabilities.

$$\rho_s(E) = \alpha + \max_{q:qEs=0} \rho_s(q)$$

Here q is an arbitrary point, not necessarily having integer coordinates. We can see how to calculate the above by substituting our expression for $\rho_s(q)$:

$$\rho_s(E) = \alpha + \max_{q:qEs=0} \max_{\hat{q}} \rho_s(\hat{q}) \exp(-||\hat{q} - q||^2)$$

$$\rho_s(E) = \alpha + \max_{\hat{q}} \max_{q:qEs=0} \rho_s(\hat{q}) \exp(-||\hat{q} - q||^2)$$

Observe that the above expression does not require us to explicitly find q. We only need the minimum distance between $\hat{q}$ and some q on the line Es. Hence,

$$\rho_s(E) = \alpha + \max_{\hat{q}} \rho_s(\hat{q}) \exp(-(\hat{q}^T l_{(E,s)})^2)$$

where the line Es is normalized as

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

where E_1 and E_2 are the first and second row of E, respectively. The final egomotion probability in the form in which it is computed is given by combining the information given by all points:

$$\rho(E) = C \prod_s [\alpha + \max_{\hat{q}} \rho_s(\hat{q}) \exp(-(\hat{q}^T l_{(E,s)})^2)]$$

Where C is chosen so that $\sum_E \rho_s(E) = 1$. This can be calculated quickly and directly from the correspondence probability distributions with no iteration. In our results, we have used $\alpha = 1$.

3.1 Egomotion Algorithm

To be totally accurate, our framework does not give a literal *answer* about what is the correct egomotion, but rather a way to calculate a distribution over the set of motions. Still, we use a simple technique to try to approximate $\arg\max_E[\rho(E)]$. Though our technique is not guaranteed to find the actual maximum, as we will discuss later, this is unlikely to make much different in performance. This is due to the fact that all parameters yielding significant probabilities tend to be contained in a very small volume of the parameter space.

First, we give our parameterization of E. We took 2 somewhat unusual parameters to represent the translation, θ and ϕ, and 3 parameters to represent the rotation r_x, r_y, and r_z. We then take $t_x = \sin(\theta)$, $t_y = \sin(\phi)$, and $t_z = \sqrt{1 - t_x^2 - t_y^2}$. If ω is a vector storing the three rotational parameters, we take R as the rotation matrix representing a rotation of angle $|\omega|$ about the unit vector $\omega/\|\omega|$. We then take the usual $E = [(t_x, t_y, t_z)^T]_\times R$.

To maximize $\rho(E)$, we first sample the parameter space equally in each of the 5 dimensions. (In the experiments given below, we used 11 points equally in each dimension, for a total of 11^5 samples.) This is followed by a gradient search initialized to each of several sample points yielding the highest probabilities. (Below we used the best 100 sample points, though in practice, a large fraction of these converge to the same answer, suggesting fewer points are necessary.)

Finally, two implementation notes: First, notice t_z is only well defined when $|\theta| + |\phi| \leq \pi/2$. Since it is more convenient to use parameters whose range is independent, in our implementation, we use parameters α and β, taking $\theta = (\pi/4)(\alpha - \beta)$ and $\phi = (\pi/4)(\alpha + \beta)$. α and β can then be allowed to vary independently from -1 to 1. Second, when maximizing $\rho(E)$, numerical properties are much improved by instead considering $\log(\rho(E))$. For simplicity, we will not discuss these issues further.

4. Experiments

4.1 'Gold Standard' Comparison

As a first experiment, we examine the relationship between the accuracy of the egomotion estimate and the number of correspondence distributions used. To give a rigorous and algorithm independent comparison, we used a least-squared epipolar minimization, based on 46 manually selected pixel-accurate correspondences. The least squared-epipolar minimization was initialized to the ground truth motion. This experiment uses two synthetic images from the well-known SOFA image database[1] (Figure 2) . While using synthetic images is unsatisfactory in some ways, the availability of the exact ground truth motion is necessary to compute errors. (The obvious way to attain the 'ground truth' for a real sequence would be to compute the motion from manually extracted correspondences, but here this very technique is being used for comparison.) We measure separately the translational and rotational error of the egomotion estimates. For the translational error, we compute the Euclidean distance $|\mathbf{t} - \hat{\mathbf{t}}|$ between the estimated focus of expansion $\mathbf{t}$, and the true focus of expansion $\hat{\mathbf{t}}$, where each is on the unit sphere. Similarly, for the rotational parameters, we calculate $|\omega - \hat{\omega}|$, where ω is a vector containing the three rotational parameters. For each size, the two algorithms were run on random subsets of that size. The mean errors for each size are shown in Figure 3. For reference, we have also included the results of running the algorithm proposed in this paper on the manually extracted matches. Here, we simply take $\rho_s(\hat{q}) = 1$ when s corresponds to $\hat{q}$ and 0 otherwise. For the algorithms using hand established matches, the means are taken over 100 random subsets of each size. For the results using correspondence probability distributions, means are taken over 25 random subsets.

It can be observed that for any given number of correspondences, the epipolar minimization will perform somewhat better than our technique on an equal number of probability distributions. Nevertheless, when using a large number of correspondences, our automatic technique is actually

[1]SOFA synthetic sequences courtesy of the Computer Vision Group, Heriot-Watt University (http://www.cee.hw.ac.uk/ mtc/sofa)

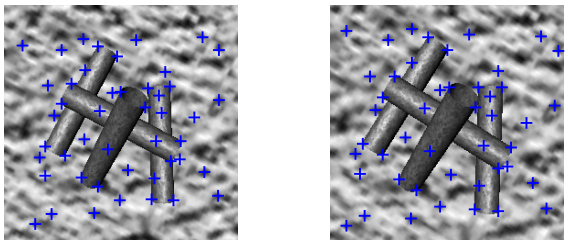

Figure 2. A synthetic sequence, with manually extracted correspondences marked.

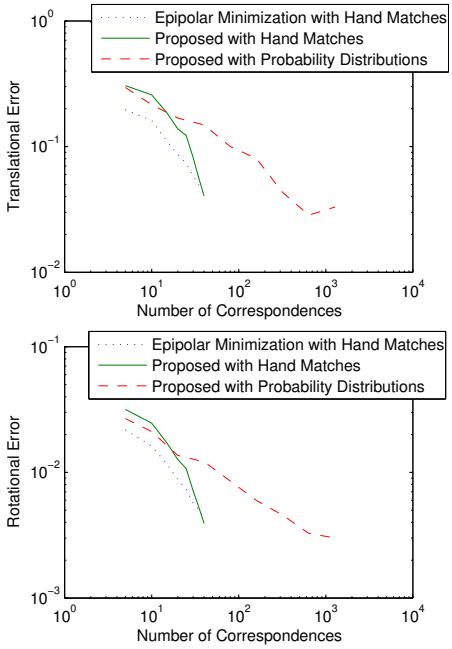

Figure 3. Mean errors for different numbers of correspondences or correspondence distributions.

able to perform comparably to this 'Gold Standard' algorithm run on manually generated pixel-accurate correspondences. The technique's success is due not to the way it processes the correspondence information, but rather to the abundance of information that is available to it.

4.2 Effects of more probability distributions

We would like to illustrate exactly how it is that the use of many distributions improves performance. To make the process easier to visualize, we fix three of the parameters to

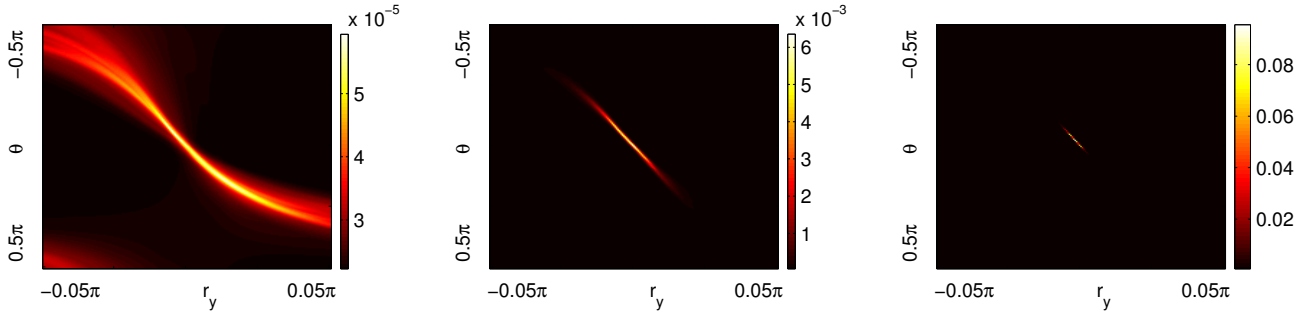

Figure 4. Three plots, each showing the computed probabilities as a function of θ and r_y. All other parameters are fixed to the ground truth. Left: 5 correspondence distributions. Center: 50 distributions. Right: 500 distributions.

the correct ground truth. It is then possible to plot $\rho(E)$ as a function of the remaining two parameters, θ and r_y, each sampled at 401 points. Figure 4 shows this for increasing numbers of input correspondence distributions. Mathematically, two correspondences known with perfect accuracy would give the exact answer. Nevertheless, a small change in the translation can be compensated by a small change in the rotation to yield similar epipolar lines. Thus, the uncertainty in the correspondence leads to an ambiguity in the motion. This ambiguity is reduced in the presence of additional correspondence information. It is for this reason we say that a better maximization of $\rho(E)$ is unlikely to significantly improve performance. Given a small number of matches, there will be a large volume in the parameter space of E all yielding similarly high probabilities. Finding a motion with a slightly higher probability can only be expected to slightly improve performance. On the other hand, as the number of input probability distributions increases, the volume of the parameter space with a high probability 'shrinks' to the correct answer. To put it in a different way, suppose we had access to a limitless number of correspondence probability distributions. In the limit, the egomotion probabilities would become $\rho(\hat{E}) = 1$ for the correct motion $\hat{E}$, and $\rho(E) = 0$ for all others. Thus any algorithm which reliably finds an E with $\rho(E)$ within some bound of the optimal one will have its error decrease towards zero as the amount of correspondence information is increased.

4.3 Egomotion in scenes with Independent Motion

As a further test of our technique, we used the well-known 'Yosemite' sequence. The clouds at the top of the images are moving independently, and nonrigidly. Figure 5. Notice that the clouds move relative to the epipolar lines, while the rest of the image does not.

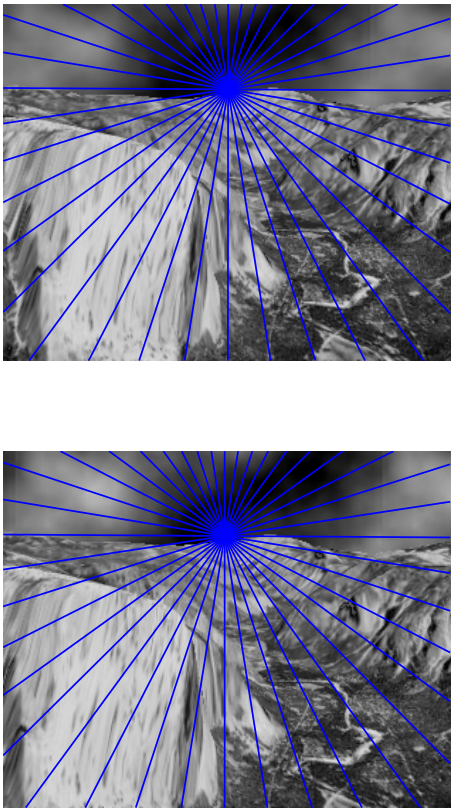

Figure 5. Two frames from the 'Yosemite' sequence, with the epipolar lines found by our method overlaid.

Finally, we captured two real sequences including independent motion. As shown in Figure 6 the rigid motion for the static background was found with high accuracy, while being undisturbed by the independently moving foreground.

5. Conclusions

Using probability distributions gives us a large amount of very robust information about correspondence. This large amount of data dramatically reduces the ambiguity in the estimation of egomotion. We have presented a technique which achieves very accurate results, even in the face of independent motion. Despite these promising results, we suspect that most aspects of our technique can be improved with further work. More accurate correspondence probability distributions could be calculated by a more rigorous examination of the imaging process. Though our simple-minded approach to maximizing $\rho(E)$ works well in practice, it would be better to have a technique with more rigorous performance bounds. Future work could also extend this framework to other problems, such as explicitly identifying which portions of the scene are independently moving.

References

[1] W. F. Clocksin. A new method for computing optical flow. In *BMVC*, 2000.

[2] D. Fleet. Disparity from local weighted phase-correlation. In *IEEE International Conference on SMC*, pages 48–46, 1994.

[3] D. J. Fleet and A. D. Jepson. Computation of component image velocity from local phase information. *Int. J. Comput. Vision*, 5(1):77–104, 1990.

[4] C. G. Harris and M. Stephens. A combined corner and edge detector. In *AVC88*, pages 147–151, 1988.

[5] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[6] A. Makadia, C. Geyer, and K. Daniilidis. Radon-based structure from motion without correspondences. In *CVPR*, 2005.

[7] D. Marr. *Vision: a computational investigation into the human representation and processing of visual information*. W. H. Freeman, San Francisco, 1982.

[8] O. Nestares, R. Navarro, J. Portilla, and A. Tabernero. Efficient spatial-domain implementation of a multiscale image representation based on gabor functions. *Journal of Electronic Imaging*, 7:166–173, 1998.

[9] J. Oliensis. A critique of structure-from-motion algorithms. *Computer Vision and Image Understanding: CVIU*, 80(2):172–214, 2000.

[10] T. D. Sanger. Stereo disparity computation using gabor filters. *Biol. Cybern.*, 59:405–418, 1988.

[11] E. P. Simoncelli, E. H. Adelson, and D. J. Heeger. Probability distributions of optical flow. In *Proc Conf on Computer Vision and Pattern Recognition*, pages 310–315, Mauii, Hawaii, 1991. IEEE Computer Society.

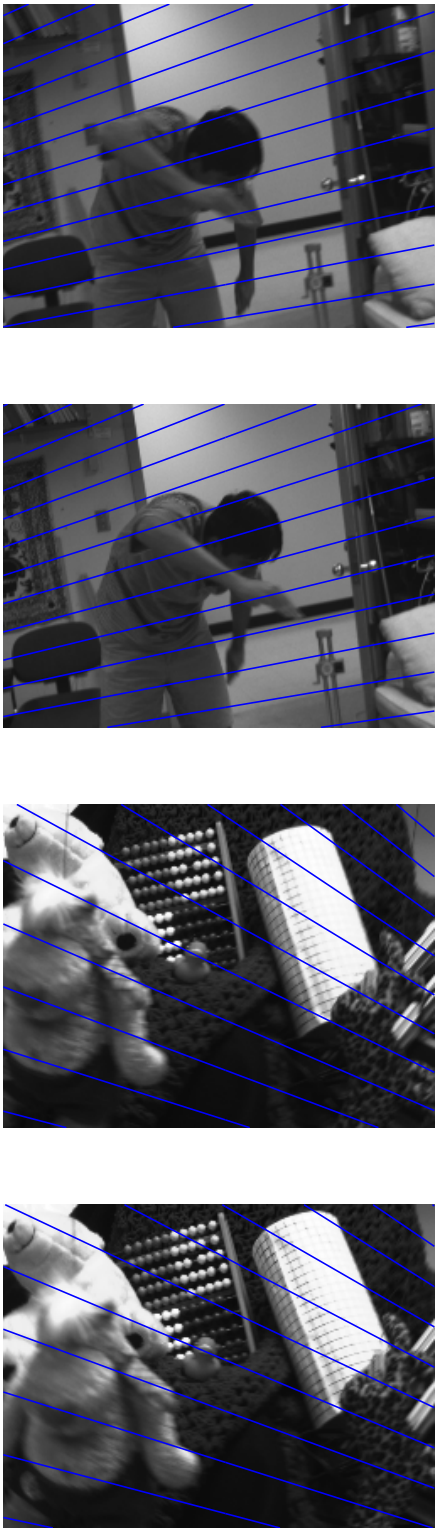

Figure 6. Frame pairs with the epipolar lines found by our method overlaid.