Feature Based Methods for Structure and Motion Estimation

P. H. S. Torr and A. Zisserman

1 Microsoft Research, One Microsoft Way, Redmond, WA 98052, USA,
  philtorr@microsoft.com
2 Dept. of Engineering Science, Oxford University, England
  az@robots.ox.ac.uk

1 Introduction

This report is a brief summary of the “feature based methods” side of the “features vs direct methods” debate. A companion paper by Irani and Anandan summarizes the “direct” side.

Many activities occur in the structure and motion literature and within this paper the two methods are contrasted applied to two problems: computation of a panoramic mosaic of the scene from a video sequence acquired by a camera rotating about its centre; and, the recovery of 3D scene structure and camera motion from a video sequence acquired by a camera undergoing general motion.

Direct methods solve two problems simultaneously: the motion of the camera and the correspondence of every pixel. Direct methods effect a global minimization using all the pixels in the image, the starting point of which is the image brightness constraint (explained in the companion paper). However we see several deficiencies with this approach: (1) the brightness constraint equation is not photometrically invariant, in that there is a tacit assumption that there is no change in intensities of the patch between consecutive images. Such a change will affect the estimated normal flow. (2) It is very difficult to adopt a probabilistic framework in that the errors in normal flow between adjacent pixels are highly correlated. Thus attempting a global minimization treating all the errors as if they were uncorrelated will lead to a biased result. (3) Normal flows can only be combined across regions of the image that have some simple parametric form (the 2D Global Motion Model) such as an affine or quadratic. It is difficult to estimate quantities such as the epipolar geometry unless the scene is mostly planar. (4) Use of only normal flow means that information is thrown away, if both directions of flow are known, why not use them? (5) Computational time is wasted by including in the
minimization a large number of pixels where no flow can be reliably estimated. Computation of a dense motion field may not even be possible in some homogeneous areas of the image. (6) The smoothing necessary in these methods causes problems at occlusion boundaries. (7) Wide baseline images cannot be tackled.

By contrast we advocate a feature based approach, and our thesis is as follows

**Thesis**: The first step of any structure and motion algorithm should be to extract features from which to estimate motion, the second step should be to use the motion to guide the dense correspondence for the non feature pixels.

Features redress many of the problems evinced by the direct methods: (1) Interest points have both geometric and photometric invariance. As shown by Schmid et al [8]. (2) The errors are uncorrelated between features so that statistical independence is a valid assumption when estimating the fundamental matrix etc. (3) There is a wide of choice of algorithms to estimate epipolar or trifocal geometry from point or line features. (4) Both components of motion estimated can be estimate at point features. (5) Areas of low information are ignored initially, resulting in a problem with far few parameters to be estimated. (6) Smoothing and the construction of a dense motion field occur only after cameras have been computed using well understood techniques such as space carving. (7) Wide baseline problems can be readily tackled.

In the next sections the feature based approach will be demonstrated and then we shall explain why it performs so well on these problems. In section 2 we will describe in detail how a feature based algorithm is used for the mosaicing case, and then in section 3 show examples of structure and motion computation.

## 2 Computing a mosaic ...

Within this section it is explained how to automatically compute a panoramic mosaic given a sequence of images acquired by a camera rotating about its centre. For this image motion corresponding image points (i.e. images of the same scene point) are related by a point-to-point map which is independent of scene structure. The map is a planar homography (also known as a plane projective transformation, or collinearity) having 8 independent parameters. This map also applies if the camera zooms (changes focal length whilst rotating), or if the scene is planar or scene relief is shallow relative to the distance from the camera.
The algorithm for computing a homography between two images is summarized in figure 1, with an example given in figure 2. Typically, there are 100s of points per image. 10-50% are matched incorrectly by similarity alone.

<table>
<thead>
<tr>
<th>Objective</th>
<th>Compute the 2D homography between two images.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Algorithm</strong></td>
<td></td>
</tr>
<tr>
<td>1. <strong>Interest points</strong>: Compute interest points in each image (e.g. Harris corners [3]).</td>
<td></td>
</tr>
<tr>
<td>2. <strong>Putative correspondences</strong>: Compute a set of interest point matches based on proximity and similarity of their intensity neighbourhood.</td>
<td></td>
</tr>
<tr>
<td>3. <strong>RANSAC robust estimation</strong>: Repeat for $N$ samples</td>
<td></td>
</tr>
<tr>
<td>(a) Select a random sample of 4 correspondences and compute the homography $H$.</td>
<td></td>
</tr>
<tr>
<td>(b) Calculate a geometric image distance error for each putative correspondence.</td>
<td></td>
</tr>
<tr>
<td>(c) Compute the number of inliers consistent with $H$ by the number of correspondences for which the distance error is less than a threshold. Choose the $H$ with the largest number of inliers.</td>
<td></td>
</tr>
<tr>
<td>4. <strong>Optimal estimation</strong>: re-estimate $H$ from all correspondences classified as inliers, by minimizing the a maximum likelihood cost function using an suitable numerical minimizer (e.g. the Levenberg-Marquardt algorithm [7]).</td>
<td></td>
</tr>
<tr>
<td>5. <strong>Guided matching</strong>: Further interest point correspondences are now determined using the estimated $H$ to define a search region about the transferred point position.</td>
<td></td>
</tr>
<tr>
<td>The last two steps can be iterated until the number of correspondences is stable.</td>
<td></td>
</tr>
</tbody>
</table>

**Fig. 1.** *Automatic estimation of a homography between two images using RANSAC.*

The two frame algorithm is extended to a sequence as follows

1. Compute interest points in each frame
2. Compute homographies between frames using interest points
3. Bundle adjust homographies and points over all frames
4. Map frames into mosaic using homographies

Details are given in [1]. The application of this algorithm to a 100 frame image sequence is illustrated in figures 3 and 4.

3 **Structure and motion algorithms**

This section gives some examples of metric reconstruction of the scene and cameras directly from the image sequence. This involves computing the
Fig. 2. Automatic computation of a homography between two images using RANSAC. The motion between views is a rotation about the camera centre so the images are exactly related by a homography. (a) (b) left and right images of Keble College, Oxford. The images are 640 x 480 pixels. (c) (d) detected corners superimposed on the images. There are approximately 500 corners on each image. The following results are superimposed on the left image: (e) 268 putative matches shown by the line linking corners, note the clear mismatches; (f) outliers — 117 of the putative matches; (g) inliers — 151 correspondences consistent with the estimated H; (h) final set of 262 correspondences after guided matching and MLE.
Fig. 3. (a) Every 10th frame of a 200 frame sequence acquired by a hand held camcorder. The motion between views is a rotation about the camera centre so the images are exactly related by a homography. (b) Result.
Fig. 4. (a) 1000 of 2500 points used in bundle adjustment, note density. (b) Every 5th frame (indicated by outline) — note lack of overlap.
cameras up to a Euclidean transformation of 3-space (auto-calibration) and a dense reconstruction of the scene. The method follows a similar method to that for homographies, this time computing fundamental matrices $F$ as described in [4]. The fundamental matrix encapsulates the epipolar geometry and hence the relative camera calibration and positions associated with two images

$$x^T F x = 0$$

(1)

A key observation to make is that this cannot be estimated from normal flow alone. This is because both components of motion can be estimated at feature points. Thus feature points provide a convenient intermediate step from input images to dense 3D reconstruction. Given the cameras, the multi-view geometry is used (e.g. epipolar geometry) to help solve for dense correspondences. An example of dense stereo reconstruction over multiple views is given in figure 5. The method is described in [6].

4 Successes of the methods

Within this section the achievements of the two methods are described and contrasted.

4.1 Current achievements with point methods

1. Point based methods can cope with severe viewing distortion, and wide baseline methods becoming available ... see figure 6.
2. They facilitate automatic estimation of $F$ (fundamental matrix) and $T$ (trifocal tensor).
3. They enable automatic computation of 3D points and cameras (SFM) over an uncalibrated video sequence. i.e. 100 of frames, Also self/auto-calibration from an uncalibrated video sequence. e.g. see figure 7.
4. The final optimization of SFM is possible with a bundle adjustment over 3D points and cameras.

4.2 How do direct methods compare with this?

1. Point to point mappings ($H$) can be estimated, but limited immunity to photometric and geometric variation.
2. It is not straightforward to write down a practical likelihood function for all pixels. Modelling of noise and statistics much more complicated, and simple assumptions of independence invalid.
Fig. 5. Metric reconstruction for general motion. The input is a sequence of images acquired by a hand held camera. The output is a 3D VRML model of cameras and scene geometry. (a)–(c) 3 views (of 6) acquired by a hand held camera. Figures courtesy of Marc Pollefeys, Reinhard Koch, and Luc Van Gool.
Fig. 6. **Wide baseline matching** The trifocal tensor is estimated using corner matches and a global homography affinity score. Five of the matched points are shown together with their corresponding epipolar lines in the second and third images. The epipolar geometry is determined from the estimated trifocal tensor.

3. **F** estimation strictly impossible from normal flow. **T** estimation possible.
4. No bundle adjustment or auto-calibration for a sequence.
   it is difficult to aggregate information over many views without moving to a discrete representation
5. Limited success with wide base line.

5 Conclusions

The purpose of this paper has not been to argue against the use of direct methods where appropriate (for instance in the mosaicing problem under small image deformations). Rather it has been to suggest that for the more general structure from motion problem, where motion, calibration, and matching (structure) are all unknowns, that the most efficient way to proceed is via the extraction of photometrically invariant features to help solve the correspondence problem. The benefit being that just a few high information features can be used to find the correct ball park of the solution. Once this is found a direct method can be used to improve the result.

It is often said (by the unlearned) of feature based methods that they only furnish a sparse representation of the scene. **This is missing the point**, feature based methods are an a way of initializing projection matrices so that a dense reconstruction method can follow. There
Fig. 7. Wilshire: 3D points and cameras for 350 frames of a helicopter shot. Cameras are shown for just the start and end frames for clarity, with the camera path plotted between. The computation method is described in [2].
is a large literature concerning methods of dense stereo given the output
of a feature matcher e.g. space carving [5, 9], dynamic programming [4],
layers [7]. The extraction of features is an intermediate step, a computa-
tional artefact that culled the useless data and afforded the use of powerful
statistical techniques such as RANSAC and bundle adjustment.

Acknowledgments

For results on image sequences: David Capel, Andrew Fitzgibbon

References

   In Proc. IEEE Conference on Computer Vision and Pattern Recognition, Santa
2. A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or
   International Conference on Computer Vision, Kerkyra, Greece, pages 307–314,
   1999.
6. M. Pollefeys, R. Koch, and L. Van Gool. Self-calibration and metric reconstruction in
   spite of varying and unknown internal camera parameters. In Proc. 6th International
   Proc. IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico,