Maintaining Multiple Motion Model Hypotheses Over Many Views to Recover Matching and Structure

Phil Torr, Andrew W. Fitzgibbon and Andrew Zisserman
Robotics Research Group, Department of Engineering Science
Oxford University, OX1 3PJ.

Abstract

In order to recover structure from images it is desirable to use many views to obtain the best possible estimates. However, whilst recovering projective structure and motion from such extended sequences problems arise that are not apparent from a general viewpoint/structure approach. Foremost amongst these are (a) maintaining image correspondences consistently through many images, and (b) identifying images, within the sequence, for which structure cannot be reliably recovery. Within this paper the use of multiple motion model hypotheses is explored as an aid to solve both of these problems.

1 Introduction

The goal of this work is to obtain 3D structure and camera projection matrices from sequences of uncalibrated images. This is used as a basis for building 3D graphical models from an image sequence; e.g. from sequences obtained from a hand-held camcorder where the motion is unlikely to be smooth, or known *a priori*.

Several systems [1, 8, 12] have now been developed which are aimed at recovering 3D graphical models. All are underpinned by the necessity to match tokens (usually interest points) successfully through image sequences with a large number of frames. Indeed it transpires that the correspondence problem is the most difficult part of structure recovery, and this is especially so when these correspondences must be maintained through many images.

It is well known that the rigidity constraint (epipolar geometry) can be used to guide matching, and this is reviewed in Section 2. Extending the idea (of guided matching) to many views equates to using structure to guide matches. However, Section 3 describes practical and theoretical limitations which apply to any projective structure and motion schemes: Firstly, sometimes the data do not contain enough information to accurately recover the epipolar geometry. In this case a simpler model must be used to guide the matching; thus the effect of using an image-image homography is explored as a guide to matching. This relation is appropriate when all the points viewed lie on a plane; or the camera is undergoing a rotation about its centre. Secondly, and even worse, it may not be possible to put all the recovered structure into the same frame in certain circumstances. This is explored further in Section 7; with possible avenues of future work given in Section 8.

2 Review: the correspondence problem over multiple views

The multiple view relations induced between images by rigid motion can be used to guide the matching [1]. These are the epipolar geometry between view pairs, represented by the fundamental matrix $\mathtt{F}$ [2, 4]; and the trifocal geometry between view triplets, represented by the trifocal tensor $\mathtt{T}$ [5–7]. Both these relations can be computed automatically from the images.

Correspondences are first determined independently between all consecutive frames as follows. An interest-point operator extracts point features ("corners") from each frame of the sequence. Correspondences are then established between pairs of frames by simultaneously estimating epipolar geometry and matches consistent with this estimated geometry. The estimation algorithm is robust to mismatches [10, 13]. This basic level of tracking is termed the **F**-Based Tracker.

A proportion of the **F**-Based tracker matches are erroneous (outliers). Many of these mismatches are removed during the simultaneous estimation of the trifocal geometry and consistent matches. This is a more powerful constraint than the epipolar because the position of a match is completely constrained, rather than just restricted to a line. Equivalently, once camera

matrices and structure are instantiated (from the trifocal tensor) structure can guide matching over any frame of the sequence.

The robust nature of the algorithm means that it is not necessary to restrict putative correspondences to nearest neighbours or even the highest cross-correlation match; as the rigidity constraint can be used to select the best match from a set of candidates. Typically the search window for candidate matches is 100 pixels, which is adequate for most image sequences. It is necessary that the scene is for the most part rigid, but small discrepancies–such as those caused by shadows and specularities–are excised automatically by the robust estimation. Given these restrictions the performance is excellent, with sequences of 100s of frames being matched automatically.

In summary, existing matches are verified using the trifocal tensor or projected structure. However, the basic mechanism for obtaining new matches is always the **F**-Based Tracker.

3 The problem of degeneracy

The strategy outlined in the last section assumes *general* camera motion and *general* position for structure. The problem is that methods developed under generality assumptions fail when these assumptions do not hold. Situations of this type are termed degenerate. In the correspondence algorithm outlined above the **F**-Based Tracker may fail for an image sequence where the camera motion is *not* general or the structure is *not* general.

The common non-general situations are of two types, both connected with planes. The first (A) is a motion degeneracy where the camera rotates about its centre with no translation. The second (B) is a structure degeneracy where all 3D points in the view are coplanar. The two cases are illustrated schematically in figure 1. In both these cases it is not possible to determine the epipolar geometry between consecutive frames: as in (A) the epipolar geometry is not defined, and in (B) it cannot be uniquely determined from image correspondences alone. However, in both cases the image correspondences are related by a homography.

In case (A) provided 3D-2D correspondences can be established between existing 3D structure and its image points in the current frame, then a camera matrix **P** can be computed from the 3D-2D correspondences. In (B) the situation is more serious because even if 3D-2D correspondences can be established in the current frame, a (projective) camera matrix cannot be computed without additional information. This is because correspondences between coplanar 3D projective structure and their images only determine a homography with 8 degrees of freedom (DOF), so that 3 of the 11 DOF of the camera matrix are undetermined. This means that if at any frame within the sequence all the viewed points arise from 3D points on plane, then the 3D structure points recovered from views seen after the plane cannot be put into the same 3D projective frame as those seen before, unless there are some correspondences between the scenes that are off the plane. This is a substantial problem that is not unique to our system but will occur in any uncalibrated structure and motion recovery algorithm.

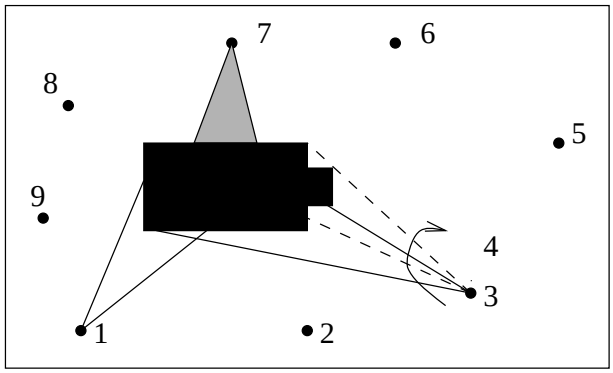


Figure 1: An illustration of how the two types of degeneracies occur. A degeneracy of type (A) occurs between frames 3 and 4 where the camera rotates about its centre. A degeneracy of type (B) occurs at frame 7 where a plane fills the field of view.

In the light of this it is clear that the ideal strategy would be to switch to homography, rather than epipolar, matching when a non-generality is reached. In both cases such homography matching maintains correspondences through the non-general frame sets. In case (A) maintaining correspondences is sufficient to completely alleviate the problem because camera matrices can then be computed for all frames and new correspondences established. Of course structure should not be instantiated from frames related by a rotation; motivating the need to detect this case. In case (B) a further strategy is required to place the correspondences in the same frame; this possibility is returned to later.

The primary issues then, which are the substance of the rest of this paper, are (i) how to detect that homography matching is necessary, i.e. an alert mechanism; (ii) if the data are consistent with a homography to distinguish whether the homography arises from case (A) or case (B) above; (iii) if type (B) how to integrate



Figure 2: Sequence including segments of pure rotation. Frames 5-9 are related to their successors by a rotation about the camera centre, whilst for the rest of the sequence there is a non-zero translation between frames.

the new frame into the old so that general matching and structure recovery can resume.

These problems are not just of theoretical interest, although they are largely ignored in theoretical developments where general motion and structure are explicitly assumed. In theory the probability of rotations exactly about the camera centre might be vanishingly small. However, in practice where relations are estimated from imperfect image measurements, the "best" (to be made more precise below) relation to use may not be that arising from general motion and structure. Especially for hand held cameras ignoring either degeneracies of type (A) or (B) means that the tracking system will often quite literally be "stopped in its tracks".

Ground truth sequence Matching using different relations is now illustrated by a 12 frame sequence, Figure 2, for which the ground truth motion type is known. The sequence was acquired under controlled conditions using a camera mounted on a mobile robot. There is general motion (non-zero translation) between frames 1 to 5; rotation about the camera centre between frames 5 to 10; and, a general motion between frames 10 to 12. The scene structure is general.

The use of two matching relations, $\mathbf{H}$ and $\mathbf{F}$, for guiding matching is now explored. There are two possibilities: either the matching relation accurately describes the situation or it does not. In the latter case over fitting or under fitting might occur as summarized in Table 1.

In Figure 3 (a) tracked matches are shown just using $\mathbf{F}$ to guide the matching pairwise between images; in (b) matches are shown using $\mathbf{H}$ and in (c) matches are shown using the correct relation at each stage. It can be seen that when $\mathbf{F}$ is used for the part of the sequence undergoing a rotational motion then there is an increase in the number of false tracks. If one is not alerted to the fact that this is a false $\mathbf{F}$ and the matches are used to instantiate new structure, the result will be many erroneous 3D points. There is also a practical problem which arises from the inevitable presence of mismatches during the estimation of $\mathbf{F}$. When image points are related by a homography they are also related by a *family* of fundamental matrices. This family can be written as $\mathbf{F} = [\mathbf{e}']_\times \mathbf{H}$. A consequence of this is that when $\mathbf{F}$ is robustly estimated in these circumstances the arbitrary choice of $\mathbf{F}$ will be determined by mismatches which are outliers to the true (homography) relation.

		Actual Motion	
		General	Rotation or Plane
Relation	$\mathbf{F}$	correct	Over fitting
	$\mathbf{H}$	Under fitting	correct

Table 1: *Possible combinations of matching relation and camera motion.*

Using $\mathbf{H}$ to guide matches throughout the sequence leads to fewer matches being extracted in the part of the sequence undergoing a general motion; as might be expected. However one should not conclude that this might always be the case: if there is a translation but parallax effects are dominated by a homography then a homography might reap more correct correspondences. An example is a looming motion (ground truth non-zero translation) for which a homography representing a zoom might be appropriate.

The number of correspondences maintained though all the images by using $\mathbf{F}$; $\mathbf{H}$ and the ground truth relation at each stage is given in table 2. Incorrect tracks tend to disappear over long sequences, as they can rarely mimic rigid motion throughout, thus all of the relations yield few false matches. Using the ground-truth relation yields a slightly better result overall compared to $\mathbf{H}$ or $\mathbf{F}$. alone.



a b c

Figure 3: Correspondences using various motion models. (a) **F** only, 89 tracks, 5 errors. (b) **H** only, 55 tracks, no errors. (c) **F** and **H**, 94 tracks, one error.

4 Automatic detection of homographies

Given that the correct image relation needs to be known to recover matches and possibly structure two strategies for automatic model selection are described and contrasted. Simply taking the model that gives lowest sum of squares of residuals will always lead to the most general model **F** being selected, as this will fit best. The first method evaluates pairs of frames and allocates either an **F** or **H** motion model. The second method maintains multiple motion hypotheses and postpones a decision on which is appropriate for a particular pair of frames until further frames have been matched.

4.1 Geometric robust information criterion

The Geometric Robust Information Criterion [11] GRIC is a robust model selection criterion that is completely general. It calculates a score function for each motion model taking into account the number of inliers, outliers, the residuals e_i, the standard deviation of the error σ, the dimension of the data r (4 for two views) the number of parameters k and dimension d of the model (**F** is a variety dimension three, **H** dimension two)

$$\text{GRIC} = \sum \rho(e_i^2) + \lambda_1 dn + \lambda_2 k \ .$$

Where $\rho(e_i^2)$ is a robust function of the residuals:

$$\rho(e^2) = \min\left(\frac{e^2}{\sigma^2}, \lambda_3(r-d)\right) \ .$$

More details are given in Torr [9], as well as typical values for λ_1, λ_2 and λ_3.

Fitting both models to each image pair and evaluating the GRIC yields the results shown in Figure 4. Over this short sequence, the success rate is 100%—the ground-truth model is selected for each pair of images. Also, the stability of the model selection may be observed in the graph of the difference $GRIC_F - GRIC_H$. This shows a sharp swing in favour of the homography model at pair $(5, 6)$, and then only small variations in the value until the reversion to the general model at pair $(10, 11)$.

4.2 Multiple hypothesis tracking

For each pair of views, both models are fitted. In order to determine the tracks chosen, sets of view pairs are taken together and for each set, the possible combinations of models are considered. For a three-view set for example, the four possibilities are FF, FH, HF and HH. Matches generated by each sub-pair of the set are combined in order to generate tracks through the set, and a decision criterion is applied in order to determine which sequence of models is retrospectively chosen to represent the set. In these experiments two criteria were applied: a simple "greedy" criterion which chooses the set with the greatest number of tracks, and a criterion which counts the number of inliers to the trifocal tensor.

4.3 Evaluation of GRIC and MHT tracks

Table 2 summarizes the results of these methods for the ground truth sequence. Correspondences were evaluated manually from the first to last frames in order to provide a ground truth set of matches on which to compare the various strategies.

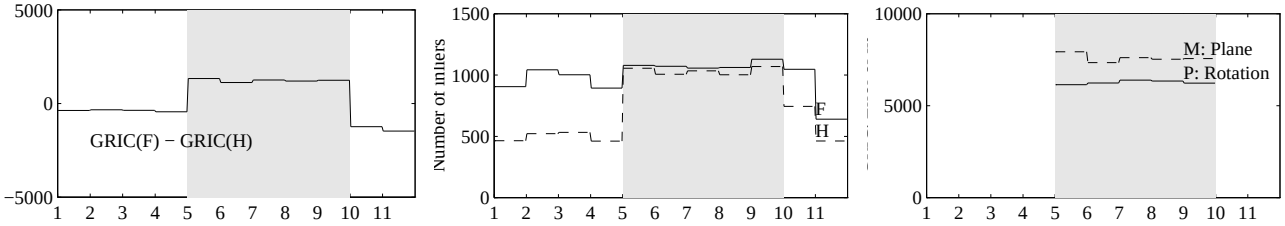


Figure 4: Selection of general (**F**) or homography (H) model. Grey denotes the region in which the camera was rotating. Left: the GRIC selects H reliably in this region. Middle: the number of inliers to H increases, but is always exceeded by **F**. Right: GRIC selects the pure rotation over the planar model within the degenerate segment.

Motion Model	Models	Correct	Wrong
F	`fffffffffffff`	89	5
Ground truth	`ffffhhhhhfff`	94	1
MHT	`ffffhfhhhhhf`	94	1
Per-track MHT	All	97	0
H	`hhhhhhhhhhhh`	55	0

Table 2: Numbers of tokens tracked through the example sequence under various motion models. Total possible tracks (determined manually) was 112.

As observed in the previous section, tracking based on the ground-truth model gives an improvement in performance. In this example, as the GRIC selects the ground-truth, it confers the same advantage. Similar performance is obtained using the greedy MHT, but in this case it is interesting to note that the ground-truth motion models were not selected. In fact, the chosen set of tracks differs in three cases from the ground truth, but the final number of matches is identical. This lends further support to the thesis that in real sequences, motion degeneracy can occur even if the camera motion contains a non-zero translation. The **Trifocal tensor** MHT produced the same set of matches; indicating that the trifocal tensor does not give an improvement over the greedy MHT.

5 Distinguishing planes from rotations

In the last section it was demonstrated that it is possible to select between homography motions and general image motions and that this should be used to improve tracking. As noted above, homographies arises in two situations: (A) the camera has undergone a rotation about its centre (B) the corresponding points all lie on a plane in 3D. In case (B) there exists a homography $\mathbf{M}$ between the (inhomogeneous) 3D structure $\mathbf{X}$ and the 2D points $\mathbf{x}$:

$$\underline{\mathbf{x}}_i = \mathbf{M} \begin{pmatrix} \underline{X} \\ \underline{Y} \\ \underline{Z} \end{pmatrix}. \qquad (1)$$

It is obviously desirable to identify this problem if it occurs, as $\mathbf{P}$ cannot be uniquely recovered, this can be done using the GRIC criterion.

Suppose that the two images viewed are related by a homography $\mathbf{H}$ and that the structure is known. To determine whether situation (A) or (B) holds the procedure is as follows; First the two 3D to 2D constraints $\mathbf{P}$ and $\mathbf{M}$ are estimated from $\mathbf{X} \leftrightarrow \mathbf{x}$ by minimizing the reprojection error, next the two GRIC scores (4.1) are evaluated. As the two images are related by a homography ($d = 2$), the GRIC differ only in the term k representing the number of degrees of freedom in the parameters, which is $k = 11$ for $\mathbf{P}$, and $k = 8$ for $\mathbf{M}$. Finally the model with lower GRIC score is deemed to be the one that holds. To be optimal the 3D points, $\mathbf{P}$ and $\mathbf{M}$ should be estimated from all views by bundle adjustment. It might be thought that measuring the residuals to a 3D plane fit would also distinguish the models, but this is impossible for a projective reconstruction as distance in $\mathcal{P}^3$ is not defined.

6 Multiple track hypotheses

Multiple hypotheses have to this point been applied to the motion models between frames. In this section MH is extended to the case in which there are multiple correspondences considered for each point in addition to the multiple motion hypotheses. This is motivated by the fact that between two views there may often be more than one candidate track for a given token, even when using just a single model. This

is most common in the presence of repeated structure, where choosing the most similar match based on intensity correlation will often lead to false matches. In a standard two-view algorithm, a "winner-take-all" strategy is necessary to gain unique matches in the absence of further information, and this strategy will necessarily lead to some of the false matches being accepted over the correct ones. As the primary idea behind MH tracking is to defer decisions until there is more information available to make them, it is natural to relax this winner-take-all condition, and accept multiple tracks for each token. By propagating these tracks into subsequent views we may determine the correct matches. In this implementation, all candidate matches, both from **F** and **H** were included, and all tracks from start to end frame generated, thereby guaranteeing the maximum possible number of correct hypotheses. However, without pruning, this also includes a large number of incorrect matches. In our system the strategy adopted is to use the multiple model hypotheses to provide a robust initialization of structure, and then excise the incorrect multi-track hypotheses using structure pruning over the entire track. On the example sequence this yields the complete set of 97 achievable matches with no errors.

7 Other sequences

Model House: In Figure 5 six images of a model house are shown; this is taken from a 32 image sequence involving a complete circumnavigation of the house. The sequence explores all the facets of the degeneracy problem as–half way round the sequence–a plane fills the view (Left Image Figure 6) and the only tracked features lie on this plane. The structure for this half of the sequence is shown in Figure 6 (b), this is put into *frame 1*. Points seen in images of the other side of the house–put in *frame 2*–cannot be placed into the same frame; until some non-planar structure of *frame 1* is seen. When this happens the two frames may be merged, and the result of this is shown in Figure 6 (c).

Mountain: Figure 7 shows the tracks from 80 images of a mountain. Although in many parts of the sequence the motion was general, the induced disparities on the image were so small that a homography **H** was sufficient to establish a good set of matches. After tracking correspondences using **H** for many frames, it was possible to the longer tracks to estimate **F** and projective structure.

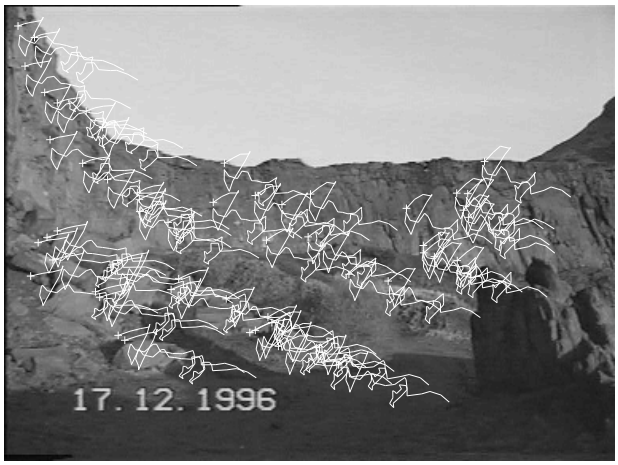


Figure 7: Tracks over 80 frames for mountain sequence.

8 Discussion and Conclusion

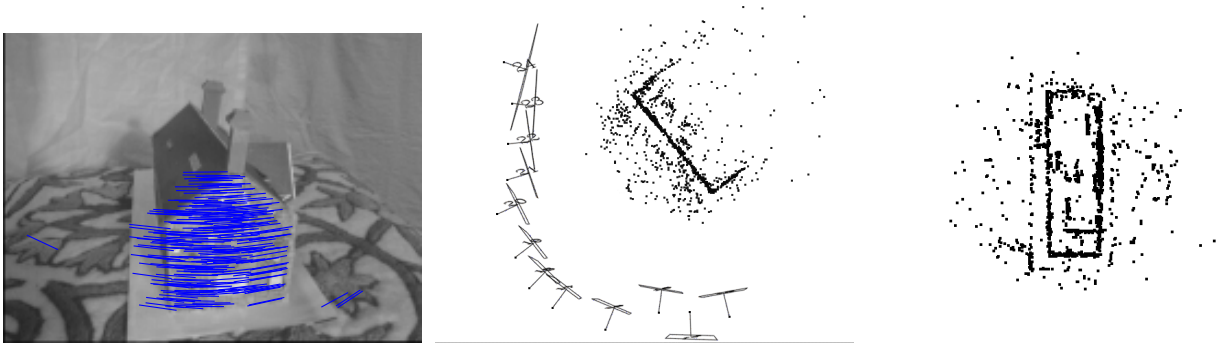
1. It has been pointed out that one strategy for coping with the type (B) structure degeneracy is to identify non-planar structure common to the two frames in order to merge them. If the reconstruction is Euclidean then coplanar points are a far less serious problem. In this case the camera calibration is known, and the motion and structure can be computed from the homography up to a *discrete* two-fold ambiguity [3].

2. If the reconstruction is affine then coplanar points also do not present the problem of a family of solutions. This can be seen as follows: the plane at infinity is effectively an additional plane available in every view; the fundamental matrix can be computed uniquely from points on two planes; so there is no ambiguity in this case.

3. The homography between images planes which arises from a pure rotation is a conjugate rotation. In this case the eigenvalues are $\{1, \exp(i\theta), \exp(-i\theta)\}$, where θ is the angle of rotation of the camera. It might be thought that this would be an effective test to distinguish homographies arising from coplanar structure from those arising from rotations. There are two reasons why this is not ideal. First, under a rotation if the internal parameters change, as they might with automatic focus, the homography is no longer a conjugate rotation. Second, the homography induced by a world plane can also be a conjugate rotation — for example the homography induced in a camera undergoing planar



Figure 5: Six images from a sequence of 32 taken of a model house rotating on a turntable.



a b c

Figure 6: (a) Tracked Matches to image 12, (b) (c) Plan view structure of the Model House shown in Figure 5.

motion by a plane perpendicular to the rotation axis.

Overall our experience with using several relations for multi-image matching has yielded some surprising conclusions. Firstly, for many images, especially if the disparity is small relative to the noise on coordinates, the homography relation **H** can yield as good a match set as the fundamental matrix, even when the underlying motion is general; although the threshold on the error used to determine inliers must be set higher for **H** than **F**. If these points are tracked across enough images they often yield sufficient disparities to recover structure stably; whereas a consecutive frame approach would be highly unstable. Secondly, the loss of 3D frame problem that occurs when a plane fills the field of view must be addressed by any system hoping to automatically deal with hundreds of unconstrained views. Within this paper we have adumbrated the problem and suggested some means of detecting and dealing with the problem.

Acknowledgements We are grateful for financial support from EU ACTS Project Vanguard. The mountain image sequence was kindly provided by the Edinburgh Virtual Environment Centre.

References

[1] P. Beardsley, P. H. S. Torr, and A. Zisserman. 3D model aquisition from extended image sequences. *ECCV4*, pages 683–695. 1996.

[2] O.D. Faugeras. What can be seen in three dimensions with an uncalibrated stereo rig? In *ECCV2*, pages 563–578. 1992.

[3] O.D. Faugeras. *Three-Dimensional Computer Vision*. The MIT Press, 1993.

[4] R. I. Hartley. Estimation of relative camera positions for uncalibrated cameras. In *ECCV2*, pages 579–587. 1992.

[5] R. I. Hartley. Lines and points in three views – a unified approach. In *ARPA IUW*, 1994.

[6] A. Shashua. Trilinearity in visual recognition by alignment. In *ECCV3*, pages 479–484, 1994.

[7] M. Spetsakis and J. Aloimonos. A multi-frame approach to visual motion perception. *IJCV*, 6:245–255, 1991.

[8] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography *IJCV*, 9(2):137–154, 1992.

[9] P. H. S. Torr. Motion segmentation and model selection. OUEL report, accepted to Philosophical Transactions of the Royal Society., 1998.

[10] P. H. S. Torr and D. W. Murray. Outlier detection and motion segmentation. In *Sensor Fusion VI*, pages 432–443. SPIE volume 2059, 1993. Boston.

[11] P.H.S. Torr. An assessment of information criteria for motion model selection. In *CVPR97*, pages 47–53, 1997.

[12] Zeller, C. *Projective, Affine and Euclidean Calibration in Compute Vision and the Application of Three Dimensional Perception*. PhD thesis, INRIA 1996.

[13] Z. Zhang, R. Deriche, O. Faugeras, and Q. T. Luong. A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *AI Jnl*, vol.78:87–119, 1994.