Computer Vision Winter Workshop 2006, Ondřej Chum, Vojtěch Franc (eds.)
Telč, Czech Republic, February 6–8
Czech Pattern Recognition Society

Surface Normal Aided Dense Reconstruction from Images

Zoltán Megyesi, Géza Kós, and Dmitry Chetverikov

Computer and Automation Research Institute
Hungarian Academy of Sciences
megyesi@sztaki.hu, kosgeza@sztaki.hu, csetverikov@sztaki.hu

Abstract

Reconstruction of 3D scenes from images is a popular task of computer vision with many applications. However, due to the inherent problems of using visual information as source, it is hard to achieve a precise reconstruction.

We discuss dense matching of surfaces in the case when the images are taken from a wide baseline camera setup. Some recent previous studies use a region growing based dense matching framework, and improve accuracy through estimating the apparent distortion by local affine transformations. In this paper we present a way of using pre-calculated calibration data to improve precision. We demonstrate that the new method produces a more accurate model.

1 Introduction

Accurate dense reconstruction from images is a challenging task of computer vision. The literature describes solutions for different viewpoint setups, ranging from the classical short baseline stereo [12, 22] to reconstruction from video [13, 17, 10] or multiple wide baseline images [21, 16, 19, 7, 11, 9].

From the setups reconstruction from short baseline images and video are well studied, but the applied methods require large overlap and similarity between frames. Most methods fail in wide baseline case due to the high distortion and occlusion rate. While these methods yield dense and reliable reconstruction, it has been pointed out in [15] and [9], that wide baseline images can be used to achieve much higher accuracy. Thus the wide baseline case cannot be ignored.

To calculate 3D depth, most methods rely on identifying different views of the same physical 3D entity. Unfortunately, often the views do not hold sufficient visual information, either because of scene properties (lack of texture, difficult view angles, lighting conditions) or the properties of the camera setup (camera distortions, image quality). Thus, the problem of finding projections of the same 3D point on the images is ill-posed and ambiguous.

As far as accuracy is concerned, the most sensitive issue is matching. Assignment of pixel correspondences between views requires the projections to be visible and distinguishable from other pixels. Since this is not the case, one must apply considerable number of geometric and visual constraints to solve the matching.

The popular modern framework for dense matching is based on global energy term minimisation [12, 16, 21]. Besides the most common *epipolar* and *brightness constancy* constraints, these methods apply different relaxation terms to reduce ambiguity, favour smooth surfaces or handle occlusions. These methods have been applied mainly to short baseline image sequences, but in [16] it has been shown that by modifying the *uniqueness* and the *ordering* constraints, the framework can also be applied to high resolution wide baseline images.

The more traditional framework assigns dense correspondences by matching pixel surroundings using an error or correlation function. These methods also exploit the above mentioned constraints, but they face problems in handling occlusions, distortions or applying smoothness constraints without proper initialisation. To cure these problems a *region growing* scheme was introduced by Lhuillier et al. [4]. This propagation framework initialises matching by reliable seed points, and spreads outwards from them. In [8] it was shown that the framework can also be applied to wide baseline images, if the correlation function is modified to handle the apparent distortion of wide baseline images. This distortion can be estimated by an affine transformation and as in [9], a reliable surface smoothness constraint can be applied during the propagation.

In this paper we intend to improve the affine dense matching method of [9] by introducing a novel surface smoothness constraint. Instead of local affine transformations we use surface normals, calculated from the apparent affine distortion and the internal camera parameters. Since the camera parameters are also needed to construct the 3D model of the scene, this last requirement is not restrictive. A number of efficient auto calibration methods are available [2, 6].

A General Reconstruction Process for Multiple Images

The reconstruction of a 3D scene from multiple images is a complex task. A general multi-step solution is presented in [14, 18, 22].

To build 3D models we first need to recover geometric information about the image planes and cameras. This requires a certain number of initial correspondences between the views. The correspondences can be assigned automatically in the short baseline case by tracking corners [13]. In the wide baseline case matching affinely invariant features gives a solution [20, 7]. If enough views are given, the cam-

eras can be self calibrated based on the initial correspondences [2, 6].

In the next step, dense matching is performed between the frames to calculate a dense depth map of the scene. This depth map can be converted to a 3D model using the previously acquired calibration data. Since the dense matching is the most time consuming and sensitive step of the reconstruction, it must be aided with as much information as we have. In the first step, both sparse correspondences and geometric information are recovered. The correspondences can be used to initialise the matching, while the geometry can be turned into constraints, like the epipolar one. Very often the images are *rectified* [2, 18] to exploit this constraint and speed up matching. In this paper we present a novel way to use the geometric information to aid the matching.

Figure 1: Rectified wide baseline image pair.

Problem Statement and Goals

In this article we give a solution to precise reconstruction from wide baseline images through an accurate dense matching method. We consider the calibration data granted, and deal with rectified image pairs. (See figure 1.) The basis of our method is the Affine Dense Matching proposed in [9].

We present a solution for extracting the surface normal of a surface patch from the affine distortion estimated between its views. The method uses some basic parameters of the camera to obtain the required geometric information. We replace the smoothness constraint used in [9] by a novel constraint based on the surface normals. We demonstrate that the new method produces more dense and accurate models.

2 Affine Dense Matching

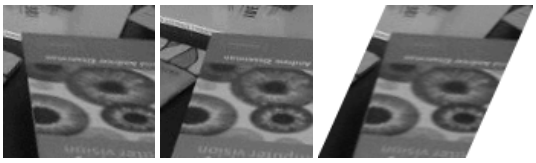

Figure 2: Left to right: Rectified sample of left image, rectified sample of right image, left sample transformed by an affine transformation to match right sample.

We consider the wide baseline case, when the images are taken from significantly different viewpoints. The two main problems of the wide baseline setup are increased distortion and occlusions. The apparent distortion can be so significant that the classical correlation based methods can fail entirely. In [8], the distortion was estimated by an affine transformation, and the estimate was used in calculating the score of a classical correlation function such as the Zero-mean Normalised Cross Correlation (ZNCC). Figure 2 shows an example. This idea can make ZNCC and other classical functions useful in dense matching under wide baseline conditions.

To reduce the computational cost, a region growing scheme is applied. As an initialisation, automatically selected textured seed points are matched by searching for the best affine transformation and disparity. Since the reliability of the seeds is crucial, inconsistent and unreliable matches (matches with low correlation score) are discarded.

The second step is the propagation of these parameters starting from the seed points towards surrounding pixels in the reference image. The propagated affine parameters and disparity are refined to get a better fit of the surface. The propagation stops when the affine transformation is not applicable any more.

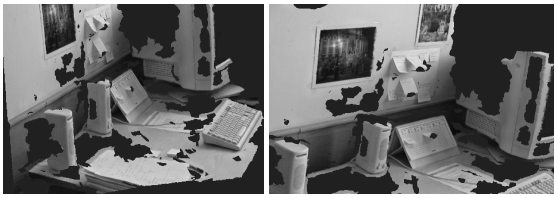

Figure 3: Result of Affine Matching, 3D representation.

In [9], the affine transformation was used also as a smoothness constraint, to control disparity search using the fact that disparity change is not independent of the propagated best affine transformation.

Papers [8, 9] were based on the connection between the facing direction (normal vector) of the 3D surfaces and the distortion in their projections, but they did not exploit the connection to the full extent. In the next section we present a way to compute the surface normals from the estimated affine transformations, clarifying the relation between them. We assume some camera calibration data to be known.

3 Normals from Affine Transformation

Estimated affine distortions have been used before in Shape-from-Texture to calculate surface normals if a reference shape is provided ([5]). In this article we use calibration data as reference for the normals.

Let us assume we have two images created by identical cameras from different viewpoints. The images are already rectified, i.e., the image planes are transformed to be coplanar, the corresponding epipolar lines coincide, and the epipoles are at infinity. We also assume that we know the (rectified) principal points (the orthogonal projections of the camera centres to the rectified viewing plane) and the camera focal distances.

Consider the projections of a 3D surface patch onto the two images and the approximate affine transformation that distorts one projection into the other one. Our goal is to estimate the normal vector in the centre of the surface patch.

3.1 Notations

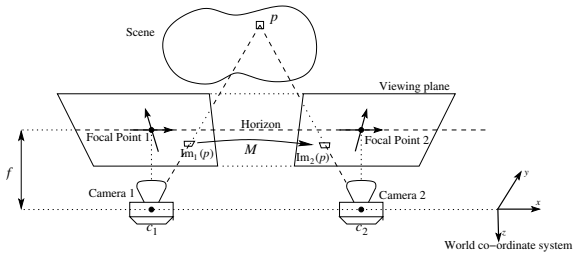

Figure 4: Notations.

Figure 4 shows our basic notations. The two rectified cameras project the view to the common viewing plane. We choose the global coordinate system such that direction x is the horizontal, y is the view-up direction and $-z$ is the common viewing direction. The origin is set so that the third coordinates of the camera centres are zero. The common focal distance of the cameras is denoted by f.

In the viewing plane we define two co-ordinate systems for the images of the two cameras. The origins are the principal points and the axes point in x and y directions, respectively. For the simpler calculations, we use homogenous coordinates for the images; the third coordinate is always $-f$.

Throughout this section we use lower indices to indicate the coordinate index of vectors, and upper indices to separate variables belonging to the first or the second image. For an arbitrary point $p = (p_1, p_2, p_3)$ of the scene, denote its two images by $\mathrm{Im}^1(p)$ and $\mathrm{Im}^2(p)$. By the projection

$$\mathrm{Im}^1(p) = \frac{f}{-p_3}(p - c^1) \text{ and } \mathrm{Im}^2(p) = \frac{f}{-p_3}(p - c^2).$$

Assume that we are observing the small neighbourhood of a point p in a smooth surface; moreover, the relation between the neighbourhoods of $\mathrm{Im}^1(p)$ and $\mathrm{Im}^2(p)$ is approximated by an affine transformation. Represent the transformation by a 3×3 matrix M; if q is in the neighbourhood of p, then $\mathrm{Im}^2(q) \approx M\mathrm{Im}^1(q)$. Since the cameras are rectified, the second co-ordinates always match and M is always of the form $\begin{pmatrix} m_1 & m_2 & m_3 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$.

3.2 Computing Normal Vector at p

Let $q = (q_1, q_2, q_3)$ be an arbitrary point in the neighbourhood of p. Then

$$c^2 - c^1 = \left(q + \frac{q_3}{f}\mathrm{Im}_2(q)\right) - \left(q + \frac{q_3}{f}\mathrm{Im}_1(q)\right) \approx$$
$$\approx \frac{q_3}{f}M\mathrm{Im}_1(q) - \frac{q_3}{f}\mathrm{Im}_1(q) = (M - I) \cdot \frac{q_3}{f}\mathrm{Im}_1(q) =$$
$$= (M - I)(c^1 - q).$$

The same holds for p as well; taking differences,

$$(M - I)(q - p) = \begin{pmatrix} m_1 - 1 & m_2 & m_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}(q - p) \approx 0.$$

Therefore, the vector $(m_1 - 1, m_2, m_3)^T$ is always perpendicular to $q - p$, so this is the normal vector of the observed surface patch.

3.3 Implementation

Using the notation of [8], the locally estimated affine transformation is a 2×2 matrix of the form $\begin{pmatrix} a_1 & a_2 \\ 0 & 1 \end{pmatrix}$. To calculate normals in that notation, consider the the corresponding 2D point pair (u, v) belonging to the 3D point p, in their original image coordinate systems. For a $\hat{p}$ in the vicinity of p,

$$A(\hat{u} - u) + u + d \approx \hat{v},$$

where $\hat{u}$ and $\hat{v}$ are the projections of $\hat{p}$, and $d = v - u$ is the disparity in u. If o^1 and o^2 denote the 2D principal points, then the matrix M can be given by $m_1 = a_1$, $m_2 = a_2$,

$$m_3 = \frac{(A(\hat{u} - o^1) - (\hat{v} - o^2))_1}{f},$$

where $(b)_a$ means the a^{th} coordinate of vector b.

To calculate affine transformation from the normals consider the normal n in the 3D point p. The affine transformation for a given corresponding pair (u, v) belonging to p can be written as $a_1 = n_1 s + 1$, $a_2 = n_2 s$, where

$$s = \frac{((o^2 - o^1) - d)_1}{fn_3 - n_1(u - o^1)_1 - n_2(u - o^1)_2}.$$

4 Using Normals in Affine Matching

In [8, 9] a region growing scheme was applied. Affine parameters were estimated for a set of seed points, and in a growing step these parameters were propagated. The affine parameters represented a *smoothness constraint* that replaced disparity search by the search for the best fitting parameters. It was assumed that on a smooth surface the best fitting affine transformation changes smoothly. The problem is, this transformation changes in a biased way. We impose the same kind of constraint that does not have the above limitation: On a smooth surface the surface normals change smoothly.

4.1 Normals and Seed Points

As shown in section 3.3, the normals can be determined from the affine transformation between the views, using the camera parameters. To start region growing, the best affine parameters have to be estimated for the seed points. As in [8], this can also be done by exhaustive search, but instead of a searching a fixed set of affine parameters, different normal directions are tried. To check whether a normal is better than another one for a corresponding pair (u, v), we calculate the affine transformation A belonging to it. A small window around u is distorted by A, and an intensity based similarity function is used to compare it with the window around v. The normal with the best score is chosen. This costly search is only done for the seed points.

4.2 Propagating Normals

In each growing step of our method the best normals are propagated. The parameters for refinement and stopping conditions are angles. The changing of affine parameters is therefore replaced by rotating the normals in coaxial circles around themselves with increasing radius. (See figure 5.)

This can be implemented efficiently by pre-calculating the rotation matrices.

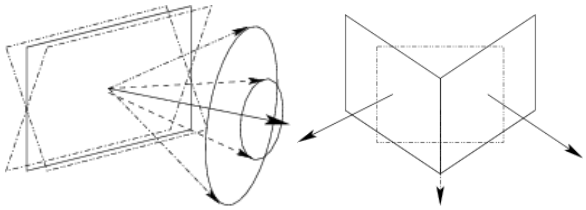

Figure 5: Left: Refinement of normal vectors. Right: Normal vectors near edges.

In the refinement step of the propagation we rotate the propagated normals by a small angle. (E.g., stepping from 1° up to 5°.) We test them by calculating the affine transformations, computing the disparity change and calculating ZNCC as in [9]. The best scoring direction will be the refined normal in the current position.

To define the stopping condition, we rotate the normal in a larger circle around itself (20–40°). If the rotated normals performed better than the refined one, it means that the surface normal changed significantly, indicating that we moved over an edge. (See illustration in figure 5.) Stopping at this conditions does not mean the pixel is not going to be part of the model: we only leave it to another seed.

5 Results and Evaluation

In this section we present some experimental results received on synthetic data, including the data used in [9]. An evaluation is also done and compared with the results of [9].

5.1 Results of Calculating Normals

The first example shows the results of normal vector calculation for a flat surface slanted by a known angle. (See figure 6.) A random texture was projected on the surface, to enable the matching. The calculated normals are shown for the seed points in figure 7.

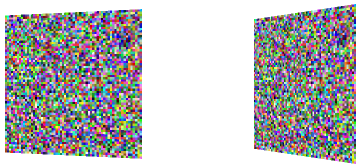

Figure 6: Slanted surface with random texture.

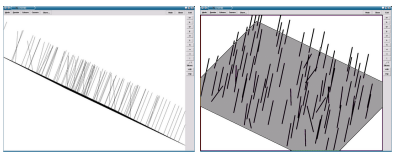

Figure 7: Calculated normals for the seed points.

5.2 Results on Semi-synthetic Datasets

In this section we present results on semi-synthetic datasets. The test objects were measured by a laser scanner, and a real texture was projected onto the model surface using the method [3]. Snapshots of the textured model were taken from different viewpoints by virtual cameras. The snapshots were created with increasing baseline width. All camera parameters were known. Our method is compared with the affine method and a classical dense matching method using ZNCC.

Shell dataset

In this test 3 image pairs were created with increasing baseline width. (See figure 8.) A sample result can be seen in figure 9.

Figure 8: Synthetic images of a ground-truth object with increasing baseline width.

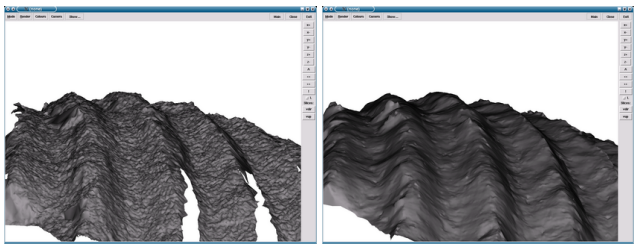

Figure 9: Reconstructed polyhedral model from the first test pair with affine method (left) and suggested method (right).

We compared the result on the synthetic shell dataset with the results of [9] using the evaluation method presented there. This evaluation method states that the precision of a reconstructed model should not consider simply the average distance from the groundtruth, but it should first separate matches found in right positions (inliers) from those in definitely wrong positions (outliers). The accuracy of the method can be measured on how dense the inlier set is, how accurate the inlier set is, and finally, what ratio of inliers to the whole set is. In the evaluation method, the inliers are automatically selected. The selection is performed using Least Median of Squares (LMedS) outlier rejection based on the distance from the groundtruth. The result of evaluation can be seen in figure 10.

One can see that the number of inliers in our method is much higher than those of the other methods. It is performing surprisingly well even under the widest baseline width. This is due to the unbiased stopping condition. The accuracy seems close to the affine solution, but in fact it is much

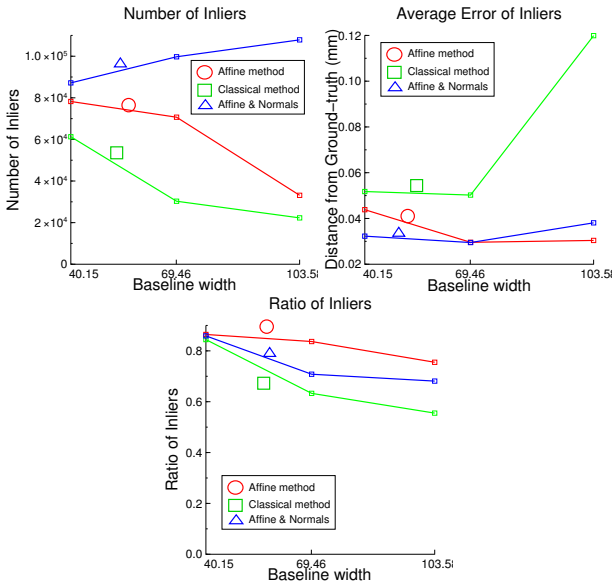

Figure 10: Evaluation of Shell results

better because of the high number of inliers. The ratio of inliers is lower than in the affine method because the affine method chooses to stop more easily.

Cat dataset

For this experiment 24 frames were created in a virtual turntable manner around the object as seen in figure 11. The view angle between frames is $15°$. The matching has been performed on adjacent frames.

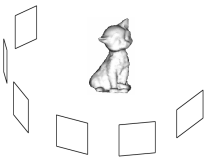

Figure 11: Virtual turntable around object

Figure 12: Consecutive frames from the Cat sequence (24 frames)

We used the same evaluation method as in the previous section, but instead of comparing methods with different baseline widths, we show the values for different pairs of frames. This way we can compare methods more reliably. It is visible that the view angle is small enough for the classical ZNCC to produce acceptable results. Figures 13 show that although the new method is slightly sparser, the improvement in accuracy is significant. The reconstructed model can be seen in figure 14.

Figure 13: Evaluation of Cat results

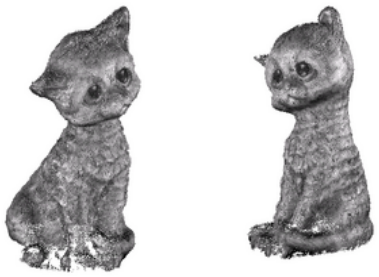

Figure 14: 3D point cloud of Cat with texture

Results on real dataset

Figure 15: 3 frames from the Monkey dataset.

We applied the algorithm to 3 distinct frames of the Monkey Dataset [1]. (See figure 15.) The dataset contains 89 shots of a difficult scene with accurate camera calibration data. The main object in the scene is covered with fur, thus it is very difficult to reconstruct correctly. However, due to the applied normal based smoothness constraint we received a smooth surface that follows the main fluctuations of the fur. The results are shown below in figure 16.

6 Conclusions

In this paper we presented a region growing based dense matching method that uses a novel and unbiased smoothing constraint. The constraint is based on the continuity of surface normals. We showed a simple way of extracting surface normals form the apparent affine distortion and camera calibration data. The method was tested on semi-synthetic groundtruth datasets, and applied on real world images with success.

We demonstrated that the method capable of accurate and dense reconstruction of difficult surfaces even under wide baseline condition.

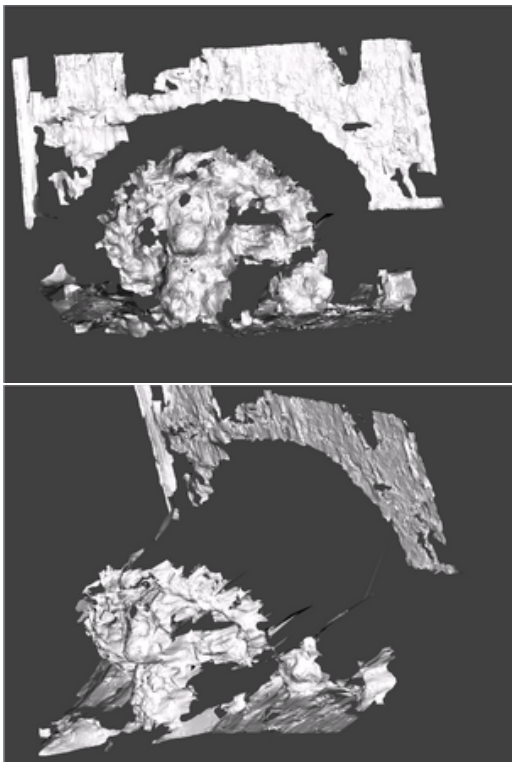

Figure 16: Triangulated 3D model of Monkey.

Acknowledgement

This work was supported by the EU Network of Excellence MUSCLE (FP6-507752) and by the Bolyai Grant of the Hungarian Academy of Sciences.

References

[1] A.W. Fitzgibbon. *Monkey Dataset*. `http://www.robots.ox.ac.uk/ awf/ibr/`.

[2] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, UK, 2000.

[3] Z. Jankó and D. Chetverikov. Photo-Consistency Based Registration of an Uncalibrated Image Pair to a 3D Surface Model Using Genetic Algorithm. In *Proc. 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, Thessaloniki, 2004.

[4] M. Lhuillier. Efficient dense matching for textured scenes using region growing. In *Proc. British Machine Vision Conf.*, pages 700–709, 1998.

[5] A.M. Loh and R. Hartley. Shape from non-homogeneous, non-stationary, anisotropic, perspective texture. In *Proc. British Machine Vision Conf.*, pages 69–78, 2005.

[6] D. Martinec and T. Pajdla. 3D Reconstruction by Fitting Low-rank Matrices with Missing Data. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 198–205, 2005.

[7] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust Wide baseline Stereo from Maximally Stable Extremal Regions. In *Proc. British Machine Vision Conference*, volume 1, pages 384–393, 2002.

[8] Z. Megyesi and D. Chetverikov. Affine propagation for surface reconstruction in wide baseline stereo. In *Proc.* 17^{th} *International Conference on Pattern Recognition*, 2004.

[9] Z. Megyesi and D. Chetverikov. Enhanced surface reconstruction from wide baseline images. In *Proc. 2nd International Symposium on 3D Data Processing, Visualization, and Transmission*, 2004.

[10] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *International Journal of Computer Vision*, pages 207–232, 2004.

[11] R. Sara. Finding the largest unambiguous component of stereo matching. In *Proc. European Conf. on Computer Vision*, volume 2, pages 900–914, 2002.

[12] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47:7–42, 2002.

[13] J. Shi and C. Tomasi. Good features to track. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Seattle*, June 1994.

[14] Milan Šonka, Václav Hlaváč, and Roger D. Boyle. *Image Processing, Analysis and Machine Vision*. PWS, Boston, USA, 1998.

[15] C. Strecha, R. Fransens, and L. Van Gool. Wide-baseline stereo from multiple views: a probabilistic account. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 2, pages 552–559, 2004.

[16] C. Strecha, T. Tuytelaars, and L. Van Gool. Dense matching of multiple wide baseline views. In *Proc. Int. Conf. on Computer Vision*, volume 2, pages 1194–1201, 2003.

[17] M. Trajković and M. Hedley. Robust recursive structure and motion recovery under affine projection. *Proc. British Machine Vision Conference*, September 1997.

[18] E. Trucco and A. Verrio. *Introductory Techniques for 3-D Computer Vision*. Prentice Hall, 1998.

[19] T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Proc* 3^{rd} *Int. Conf. on Visual Information Systems*, pages 493–500, 1999.

[20] T. Tuytelaars and L. Van Gool. Wide baseline stereo based on local, affinely invariant regions. In *Proc. British Machine Vision Conf.*, pages 412–422, 2000.

[21] G. Zeng, S. Paris, L. Quan, and M. Lhuillier. Surface Reconstruction by Propagating 3D Stereo Data in Multiple 2D Images. In *Proc. European Conf. on Computer Vision*, pages 163–174, 2004.

[22] 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. Technical Report 2273, INRIA, 1994.