MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Robust Semi-Dense Matching across Uncalibrated and Widely Separated Views

Benjamin ALBOUY[1], Sylvie TREUILLET[2], Yves LUCAS[3]
[1]LAIC, Univ. of Clermont-Fd, France, benjamin.albouy@iut.u-clermont1.fr
[2]LASMEA, Univ. of Clermont-Fd, France, sylvie.treuillet@lasmea.univ-bpclermont.fr
[3]LVR, Univ. of Bourges, France, yves.lucas@bourges.univ-orleans.fr

Abstract

This paper proposes an iterative framework to obtain a semi-dense matching from two wide-baseline images, when camera parameters and positions are unknown. The matching process is computed in three steps. First, a small number of robust correspondences is selected with SIFT descriptor. Then, this initial set is multiplied by iteratively reinforcing the epipolar constraint on rectified images during the estimation of the fundamental matrix. At a final refinement, point correspondences are boosted by applying a local affine transform in the image. The method has been tested on several pairs of images. Between three to four thousands points are paired on medium resolution images (1024x768 pixels) by this matching method: a sufficient number for accurate volumetric measurement.

1. Introduction

Accurate 3D model reconstruction from a set of uncalibrated images is a challenging task in computer vision. It is well known that triangulation accuracy is better in a wide-baseline setup with large vergency. But the performance of the reconstruction relies on the efficiency of the matching and large camera movements cause unfortunately fewer matches. Most of the works overcome this paradox by using a large sequence of images taken from all around an object [1]. The visibility of salient points in several consecutive views is used to glue partial reconstruction in the optimization process of bundle adjustment. Due to the overlap between consecutive frames, the matching is so easier to realize gradually between wide-baseline views. In a general way, correspondences of salient points in consecutive image pairs or triples are inferred using a similarity measure and spurious matches are removed by applying the epipolar constraint. By this way, several hundreds of correct matches collected in each subset are accumulated during the sequence. However, high-quality reconstructions may be provided by bundle adjustment convergence at the only condition that all parts of the object should have a roughly equal representation in image data, so that the images were regularly spaced all around the scene.

But, in our application, such capture setup is not possible: we focus on providing an accurate reconstruction of skin wound from only two or three views captured with a hand-held digital camera in a clinical environment [2]. This way of capture let a great freedom to the clinician to adapt to all pathologies. In such a context, it is not clear how well the above methods would work with a very limited set of wide-baseline images. Considering multiple widely separated views, the most difficult problem is to infer a large enough number of matches. Unfortunately, most previous methods of dense matching developed for calibrated short base-line stereo [3] fail due to large changes in the images.

During last years, a consequent research effort was related to reliable local descriptors for sparse matching in spite of geometric and photometric distortions [4,5,6,7,8]. The last one, named SIFT, has been shown to be one of the most efficient [9]. Nevertheless, finding numerous correspondences in wide separately images remains one of the main challenges in computer vision. Even if sparse matching is sufficient to retrieve epipolar geometry, a denser mapping is needed for visualization purpose or precise volumetric evaluation. To obtain denser sparse matching, Lhuillier proposes a propagation framework starting from a set of seed points [10]. The principle is similar to a region growing algorithm based on the correlation score under a smoothness constraint of disparity gradient. Propagation framework was firstly introduced by Otto for terrain images [11]. Megyesi extends the region growing to wide-baseline images by handling local affine distortion [12]. They search the best correlation score within the affine parameter domain, reduced to only two parameters by the epipolar constraint. The method is applied to grey-scale images rectified beforehand with precision by manual or automatic initialization. Same authors improved their method by exploiting the normals on the surface assuming camera calibration [13].

The problem we address in this paper is a global framework for finding a large number of robust correspondences across uncalibrated and widely separated colour views (Figure 1). Previous dense methods could not be adapted to such case. The proposed method is based on an iterative process to boost the number of robust matches on the rectified images during the refinement of epipolar geometry, followed by a semi-dense multiplication based on local affine transformations. The paper is organized as follows. The iterative step of the method is first presented. Then, the semi-dense multiplication is described in section 3. Experimental results are given in section 4. Finally, some concluding remarks and future works finish this paper.

2. Iterative Process

The iterative matching process is computed in two steps. Firstly, a small number of robust correspondences are selected to compute a rough estimate of the fundamental matrix *F.* Then, a robust detection of a large number of point correspondences is applied by iteratively reinforcing the epipolar constraint on rectified images.

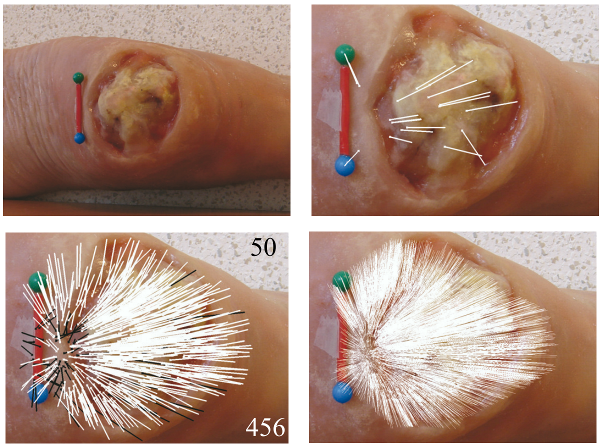

Figure 1. Two widely separated images (top) of 1024x768 pixels, initial SIFT-based matches (top right), good correspondences *in white*, outliers *in black* at first iteration (bottom left) and after the final refinement (bottom right).

2.1 Robust initialization

Matching process starts with finding some robust correspondences with a *Winner Takes All* strategy and the SIFT descriptor [8]. All initial cross-checked sparse matches are sorted by decreasing pairing score and the less ambiguous are preserved (score < 0,6). These initial matches (Figure 1 – top right) are used to compute a first fundamental matrix estimate. A standard solution to reject spurious correspondences is to use of a robust technique to jointly estimate the fundamental matrix with rejection of outliers [14] based on random sampling consensus (RANSAC). The best result is given by combining the robust Least Median of Square method (LMedS) to reject mismatches followed by an orthogonal least means square optimization on the inliers set [15].

2.2 Iteratively reinforcing epipolar constraint

To feed the iterative selection of inliers, several thousands of salient points are detected by the extension of the luminance Harris corner detector to colour proposed by [4]. New matches are generated by applying the epipolar geometry encoded in the initial estimate of the fundamental matrix. After image rectification, the matching is reduced to the canonical case where corresponding visible points are on the same row. We applied the general method of rectification by unrolling the image around the epipole in polar coordinates. This method gives good results even if the epipoles are inside the images.

The combined process of estimating F and selecting inliers is generally made once only. The usual matching algorithm reduces the search of the best correlation score to a narrow rectangular band along epipolar line. The dimensions of the rectangular band are usually fixed. The length represents the disparity limit and a narrow width is intended to mitigate the non perfect epipolar geometry. The Zero-mean Sum of Absolute Difference (ZSAD) correlation score is applied within a circular window of 31x31 pixels. The unicity is preserved by a symmetric *Winner Takes All* strategy, but some ambiguities remain if close scores are detected. Such mismatches may be rejected by considering the difference between maxima, other additional constraints, or relaxation. We choose to apply neighbourhood constraint by checking the local coherence of the pairings carried out in a zone of 50 pixels: a tolerance of 25% is authorized on the norm of matching vectors and a maximum variation of $\pi/10$ on the argument. With such an algorithm, several hundred correct matches are selected (Figure 1 - bottom left).

We propose an iterative evolution of the above method by progressively reinforcing the epipolar constraint on rectified images. The usual fixed bandwidth of the rectangular search area is replaced by a linear decreasing width during iterations. Reducing the search band improves the performance of the existing method. The epipolar geometry is then more and more efficient at each step with two benefits: more mismatches will be rejected and the cross picking of good matches is easier on a set of fewer candidates. The number of good correspondences is so boosted by this improvement: 565 matches are selected after seven iterations on the same image pair presented in Figure 1, namely 25% more than after one pass. We tested the efficiency of this decreasing band on several image pairs. Results are presented in section 4.

At the end of this iterative process, a last check is carried out on the 3D reconstructed points. An Euclidian reconstruction is realized starting from the points correspondences and the fundamental matrix based on self calibration [16]. The image formation is based on pin-hole model with simplifying assumptions: squared pixels, no bias on principal point, and same zooming factor, so that the unique unknown camera parameter is the focal length. Next, a bundle adjustment is designed to optimize the reconstructed points, the focal length and the camera positions. Finally, the points which violate the assumption of a smooth 3D surface (negative or positive peak) are considered as not reliable correspondences and removed from the 3D surface. A 3D point is rejected if a difference of more 10% is detected with the median distance calculated on the neighbours.

3. Semi-Dense Affine Correlation

The above iterative process provides a limited set of sparse consistent correspondences (typically around four or five hundreds pairs). The key observation is that a greater number of correct matches gives a more reliable epipolar geometry estimate and a more accurate reconstruction. Obtaining a denser mapping of the reconstructed surface would be advantageous with the volume estimate. A classical interpolation of the surface would only create additional 3D points without improving the accuracy. Otherwise, to use a correlation score reaches its limits on rectified images of so different viewpoints.

To multiply the point correspondences, the proposed method consists in applying some local deformations in the images after a triangulation in order to support additional correspondences within homologous triangles (Figure 2). A classical Delaunay triangulation is computed on the inliers in the first image (P_1,P_2,P_3), and transposed in the second one (P'_1,P'_2,P'_3). If the object can be assumed to be locally planar in triangular sections, projections in the two images are simply warped under homographic deformation H.

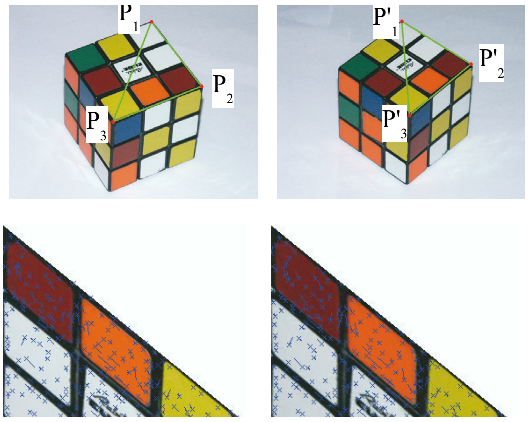

Figure 2. Triangulation and affine transformation of paired triangles for boosting matches based on correlation score controlled by a relaxation algorithm.

Table 1. Number of correspondences preserved at the end of each stage.

Number of inliers	Initial matches	Iterative matching	Final refinement
Pair 1	44	759	2962
Pair 2	40	889	3031
Pair 3	26	490	1733
Pair 4	19	565	4501

If the plane assumption is not strictly checked, the real candidate would be close to the one transformed by *H*. The position error is related to the flatness error and remains weak if the partitioning of the image is sufficient. The calculation of *H* may be based on the 3D model obtained previously. However, we choose to estimate the homographic deformation *H* by the nearest affine one. The template affine deformation is directly computed from the three paired vertices. This method is more general because it can be used even if the reconstructed vertices are not available. Considering affine deformations instead of perspectives ones causes small errors, as much smaller than the considered triangular facet of the object is flat and parallel to the images plans.

After transformation, the salient points extracted inside each warped triangular region are then more easily matched with a correlation score controlled by the relaxation algorithm proposed by Zhang (Figure 2). The correlation score is pondered by a Gaussian function (σ=5) to support the closest matches i.e. the 3D points which belong to the average plane. An example of result on real image is presented in Figure 1 (bottom right). With this final refinement, the number of correspondences increases from 565 to 4501 in this image. This greater number of matches allows a more accurate reconstruction, once again optimized by a bundle adjustment.

4. Results

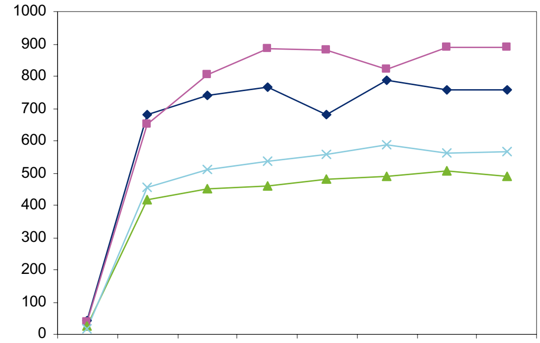

Figure 3. Evolution of the number of correspondences during iterative process.

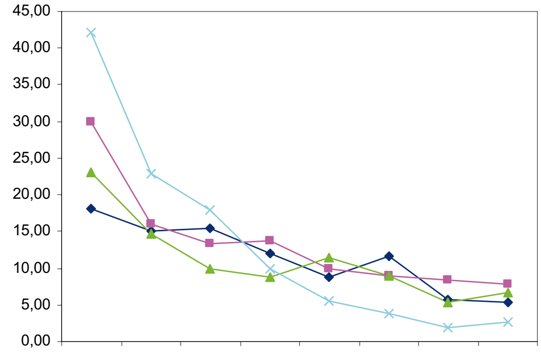

Figure 4: Evolution of the percentage of outliers during iterative process.

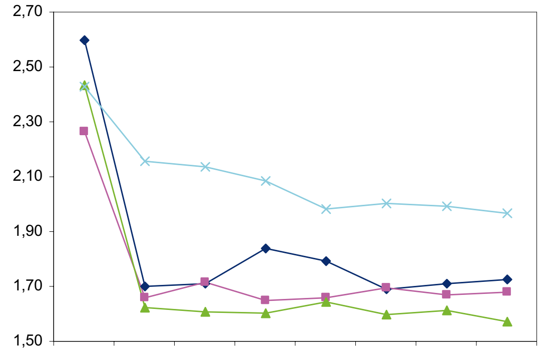

Figure 5. Evolution of the average distance to the epipolar line *in pixels*.

The iterative matching method has been tested on four pairs of medium resolution colour images (1024x768 pixels) recorded in JPEG format (high quality). Images have been captured using a standard digital camera Sony DSC-H1 with unknown focal length under widely separated points of views: the angle between optical axes varies from 23 to 47 degrees and the distance with the object between 42 and 68 centimeters. The effectiveness of the method is measured by examining the evolution during iterations of three quality standards: the number of correspondences (Figure 3), the percentage of outliers rejected by the robust estimation of *F* (Figure 4), and the average distance to the epipolar line (Figure 5).

The results show clearly that the number of correct correspondences is multiplied by progressively reinforcing the epipolar constraint during seven iterations. The number of possible correspondences is limited by the two thousand points initially provided in input of the process. The average distance to the epipolar line is bounded by the bandwidth considered at each stage. In this experiment, bandwidth has been initialized to 35 pixels and linearly decreased during seven iterations to 5 pixels. On figure 4, we can see that the matching quality is improved during the iterations, by rejecting a smaller portion of mismatches. Table 1 synthesizes the number of correspondences preserved at the end of each stage. This

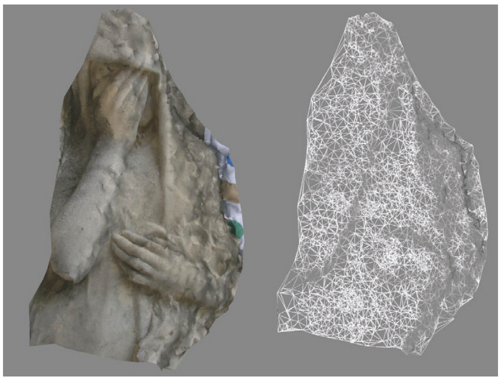

Figure 6. Textured and STL models of art work obtained from only two uncalibrated views.

process is not time consuming: around 2 minutes on a classical personal computer (Pentium IV - 3 GHz).

The proposed method has also been tested on other images extracted from the base proposed in [9]: depending on images, a multiplying factor between four and thirty is observed in relation to the number of correspondences at exit of the iterative process. This semi-dense matching offers denser mapping for nice visualization purpose (Figure 6) and a sufficient resolution for accurate volumetric measurement even from two views only. The precision of the reconstructed 3D model has been evaluated by comparison with the ground-truth provided by an industrial 3D laser scanner. The Root Mean Square error on 3D points is less than one millimeter, i.e. a relative error around 1,3% reported to the model dimensions [17].

5. Conclusion

As far as the accuracy of the inferred 3D model is concerned, the most sensitive issue is matching in images. Due to the image difficulties, the problem of dense matching from widely separated images has not been so largely addressed in the literature than in the short baseline case. However, wide baseline stereo has many advantages like a more accurate reconstruction from a smaller number of images.

Two major improvements have been introduced in the usual reconstruction chain to provide a semi-dense matching. First, an iterative process combines a progressive reinforcing of the epipolar constraint with the fundamental matrix convergence for a more robust rejection of mismatches. Secondly, a local mapping of triangles allows correcting the affine distortions to obtain a denser matching correlation. Increasing the number of robust matches results in higher accuracy and stability of the 3D model. Without providing a matching as dense as the proposed by [10], the method presented in this paper is rather simple but succeeds in providing a semi dense matching on more widely separated image pair. This density has been proved to offer a sufficient resolution for accurate volume estimation. Future works will consider images of higher resolution and address the problem of occluded areas.

Acknowledgments This work is supported by the European Social Funds and the French research ministry.

References

[1] M. Pollefeys and al.: "Visual Modeling with a Hand-Held Camera," *Int. Journal of Computer Vision*, vol.53, no.3, pp.207-232, 2004.

[2] B. Albouy, S. Treuillet, Y. Lucas, J.C. Pichaud, "Volume Estimation from Uncalibrated Views Applied to Wound Measurement," in *Int. Conference on Image Analysis and Processing*, pp.945-952, 2005.

[3] R. Scharstein and R. Szeliski: "A Taxonomy and Evaluation of Dense Two-Frame Stereo Correspondance Algorithms," *Int. Journal of Computer Vision*, vol.47, pp.7-42, 2002

[4] P. Montesinos, V. Gouet and R. Deriche: "Differential Invariants for Color Images," in *Int. Conference on Pattern Recognition*, pp.838-840, 1998.

[5] T. Tuytelaars and L. Van Gool: "Wide Baseline Stereo Matching based on Local, Affinity Invariant Regions", in *British Machine Vision Conference*, Bristol (UK), 2000.

[6] F. Schaffalitzky and A. Zisserman: "Viewpoint Invariant Texture Matching and Wide Baseline Stereo", in *Int. Conference on Computer Vision*, Canada, pp.636-643, 2001.

[7] J. Matas, and al.: "Robust Wide Baseline Stereo from Maximally Stable Extremal Regions", in *British Machine Vision Conference*, Cardif (UK), pp.384-393, 2002.

[8] D. Lowe: "Distinctive Image Features from Scale-Invariant Keypoints," *Int. Journal of Computer Vision*, vol.60, no.9, pp.91-110, 2004.

[9] K. Mikolajcyk and C. Schmid.: "Performance Evaluation of Local Descriptors," *IEEE Trans. on PAMI*, vol.27, no.10, pp.1615-1630, 2005.

[10] M. Lhuillier and L. Quan : "Match Propagation for Image-Based Modelling and Rendering," *IEEE Trans. on PAMI*, vol.24, no.8, pp.1140-1146, 2002.

[11] G. Otto and T. Chau: "A Region-Growing for Matching of Terrain Images", *Image and vision Computing*, vol.7, no.2, pp.3-94, 1989.

[12] Z. Megyesi and D. Chetverikov: "Affine Propagation for Surface Reconstruction in Wide Baseline Stereo", in *Int. Conference on Pattern Recognition*, 2004.

[13] Z. Megyesi, G. Kos and D. Chetverikov: "Surface Normal Aided Dense Reconstruction from Images", in *Computer Vision Winter Workshop,* 2006.

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

[15] B. Albouy, S. Treuillet, Y. Lucas, D. Birov, "*Fundamental Matrix Estimation Revisited through a Global 3D Reconstruction Framework*," in *Advanced Concepts for Intelligent Vision Systems*, Brussels, Belgium, pp.185-192, 2004.

[16] P. Sturm, Z.L. Cheng, P.C.Y. Chen and A.N. Poo: "Focal Length Calibration from Two Views: Method and Analysis of Singular Cases," *Computer Vision and Image Understanding*, vol.99, no.1, pp.58-95, 2005.

[17] B. Albouy, E. Koenig, S. Treuillet, Y. Lucas,, "Accurate Structure Measurement from Two Uncalibrated Views," *Lecture Notes on Computer Science*, Vol. ACIVS, Antwerp, Belgium, pp.1111-1121, 2006.