EPSC Abstracts,
Vol. 3, EPSC2008-A-00128, 2008
European Planetary Science Congress, © Author(s) 2008

Camera Trajectory from Wide Baseline Images

M. Havlena, A. Torii and T. Pajdla

Center for Machine Perception, Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University in Prague, Karlovo náměstí 13, 121 35 Prague 2, Czech Republic {havlem1, torii, pajdla}@cmp.felk.cvut.cz

ABSTRACT

Camera trajectory estimation, which is closely related to the structure from motion computation, is one of the fundamental tasks in computer vision. Reliable camera trajectory estimation plays an important role in 3D reconstruction [1], self localization [2], and object recognition [3]. There are essential issues for a reliable camera trajectory estimation, for instance, choice of the camera and its geometric projection model, camera calibration, image feature detection and description, and robust 3D structure computation.

Most of approaches [4, 5, 6] rely on classical perspective cameras because of the simplicity of their projection models and ease of their calibration. However, classical perspective cameras offer only a limited field of view, and thus occlusions and sharp camera turns may cause that consecutive frames look completely different when the baseline becomes longer. This makes the image feature matching very difficult (or impossible) and the camera trajectory estimation fails under such conditions. These problems can be avoided if omnidirectional cameras, e.g. a fish-eye lens convertor [7], are used. The hardware which we are using in practice is a combination of Nikon FC-E9 mounted via a mechanical adaptor onto a Kyocera Finecam M410R digital camera. Nikon FC-E9 is a megapixel omnidirectional add-on convertor with 180° view angle which provides images of photographic quality. Kyocera Finecam M410R delivers 2272×1704 images at 3 frames per second. The resulting combination yields a circular view of diameter 1600 pixels in the image.

Since consecutive frames of the omnidirectional camera often share a common region in 3D space, the image feature matching is often feasible. On the other hand, the calibration of these cameras is non-trivial and is crucial for the accuracy of the resulting 3D reconstruction. We calibrate omnidirectional cameras off-line using the state-of-the-art technique [8] and Mičušík's two-parameter model [7], that links the radius of the image point r to the angle θ of its corresponding rays w.r.t. the optical axis as $\theta = \frac{ar}{1+br^2}$. After a successful calibration, we know the correspondence of the image points to the 3D optical rays in the coordinate system of the camera. The following steps aim at finding the transformation between the camera and the world coordinate systems, i.e. the pose of the camera in the 3D world, using 2D image matches.

For computing 3D structure, we construct a set of tentative matches detecting different affine covariant feature regions including MSER [9], Harris Affine, and Hessian Affine [10] in acquired images. These features are alternative to popular SIFT features [11] and work comparably in our situation. Parameters of the detectors are chosen to limit the number of regions to 1-2 thousands per image. The detected regions are assigned local affine frames (LAF) [12] and transformed into standard positions w.r.t. their LAFs. Discrete Cosine Descriptors [13] are computed for each region in the standard position. Finally, mutual distances of all regions in one image and all regions in the other image are computed as the Euclidean distances of their descriptors and tentative matches are constructed by selecting the mutually closest pairs.

Opposed to the methods using short baseline images [6], simpler image features which are not affine covariant cannot be used because the view point can change a lot between consecutive frames. Furthermore, feature matching has to be performed on the whole frame because no assumptions on the proximity of the consecutive projections can be made for wide baseline images. This is making the feature detection, description, and matching much more time-consuming than it is for short baseline images and limits the usage to low frame rate sequences when operating in real-time.

Robust 3D structure can be computed by RANSAC [14] which searches for the largest subset of the set of tentative matches which is, within a predefined threshold ε, consistent with an epipolar geometry [1]. We use ordered sampling as suggested in [15] to draw 5-tuples from the list of tentative matches ordered ascendingly by the distance of their descriptors which may help to reduce the number of samples in RANSAC. From each 5-tuple, relative orientation is computed by solving the 5-point minimal relative orientation problem for calibrated cameras [16, 17].

Often, there are more models which are supported by a large number of matches. Thus the chance that the correct model, even if it has the largest support, will be found by running a single RANSAC is small. Work [18] suggested to generate models by randomized sampling as in RANSAC but to use soft (kernel) voting for a parameter instead of looking for the maximal support. The best model is then selected as the one with the parameter closest to the maximum in the accumulator space. In our case, we vote

in a two-dimensional accumulator for the estimated camera motion direction. However, unlike in [18], we do not cast votes directly by each sampled epipolar geometry but by the best epipolar geometries recovered by ordered sampling of RANSAC. With our technique, we could go up to the $98.5\ \%$ contamination of mismatches with comparable effort as simple RANSAC does for the contamination by $84\ \%$. The relative camera orientation with the motion direction closest to the maximum in the voting space is finally selected.

As already mentioned in the first paragraph, the use of camera trajectory estimates is quite wide. In [19] we have introduced a technique for measuring the size of camera translation relatively to the observed scene which uses the dominant apical angle computed at the reconstructed scene points and is robust against mismatches. The experiments demonstrated that the measure can be used to improve the robustness of camera path computation and object recognition for methods which use a geometric, e.g. the ground plane, constraint such as does [3] for the detection of pedestrians. Using the camera trajectories, perspective cutouts with stabilized horizon are constructed and an arbitrary object recognition routine designed to work with images acquired by perspective cameras can be used without any further modifications.

Acknowledgement

This work has been supported by the European IST Programme Project FP7-218814 PROVISG and by the project MSM6840770038 DMCM III. This paper only reflects the authors' views and funding agencies are not liable for any use that may be made of the information contained herein.

References

[1] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, second edition, 2003.

[2] T. Goedemé, M. Nuttin, T. Tuytelaars, and L. Van Gool, "Omnidirectional vision based topological navigation," *IJCV*, vol. 74, no. 3, pp. 219–236, September 2007.

[3] B. Leibe, N. Cornelis, K. Cornelis, and L. Van Gool, "Dynamic 3d scene analysis from a moving vehicle," in *CVPR 2007*, Minneapolis, MN, USA, 2007.

[4] 2d3 Boujou, "http://www.boujou.com," 2001.

[5] A. Akbarzadeh et al., "Towards urban 3d reconstruction from video," in *3DPVT*, May 2006, Invited paper.

[6] N. Cornelis, K. Cornelis, and L. Van Gool, "Fast compact city modeling for navigation pre-visualization," in *CVPR 2006*, 2006, pp. II:1339–1344.

[7] B. Mičušík and T. Pajdla, "Structure from motion with wide circular field of view cameras," *IEEE Trans. PAMI*, vol. 28, no. 7, pp. 1135–1149, July 2006.

[8] H. Bakstein and T. Pajdla, "Panoramic mosaicing with a 180° field of view lens," in *Proc. IEEE Workshop on Omnidirectional Vision*, 2002, pp. 60–67.

[9] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide-baseline stereo from maximally stable extremal regions," *Image and Vision Computing*, vol. 22, no. 10, pp. 761–767, 2004.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool, "A comparison of affine region detectors," *International Journal of Computer Vision*, vol. 65, no. 1-2, pp. 43–72, 2005.

[11] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] Š. Obdržálek and J. Matas, "Object recognition using local affine frames on distinguished regions," in *BMVC 2002*, London, UK, 2002, vol. 1, pp. 113–122.

[13] Š. Obdržálek and J. Matas, "Image retrieval using local compact dct-based representation," in *DAGM 2003*, Berlin, Germany, 2003, number 2781 in LNCS, pp. 490–497.

[14] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Comm. ACM*, vol. 24, no. 6, pp. 381–395, June 1981.

[15] O. Chum and J. Matas, "Matching with PROSAC - progressive sample consensus," in *CVPR 2005*, Los Alamitos, USA, 2005, vol. 1, pp. 220–226.

[16] D. Nistér, "An efficient solution to the five-point relative pose problem," *IEEE Trans. PAMI*, vol. 26, no. 6, pp. 756–770, June 2004.

[17] H. Stewénius, *Gröbner Basis Methods for Minimal Problems in Computer Vision*, Ph.D. thesis, Centre for Mathematical Sciences LTH, Lund University, Sweden, 2005.

[18] H. Li and R. Hartley, "A non-iterative method for correcting lens distortion from nine point correspondences," in *OMNIVIS 2005*, 2005.

[19] A. Torii, M. Havlena, T. Pajdla, and B. Leibe, "Measuring camera translation by the dominant apical angle," in *CVPR 2008*, Anchorage, AK, USA, 2008.