

Effective Structure-from-Motion for Hybrid Camera Systems

Yalin Bastanlar, Alptekin Temizel, Yasemin Yardimci
Informatics Institute
Middle East Technical University
Ankara, Turkey
 {yalinb,atemizel,yardimy}@ii.metu.edu.tr

Peter Sturm
INRIA Rhone-Alpes and
Laboratoire Jean Kuntzmann
Grenoble, France
 peter.sturm@inrialpes.fr

Abstract—We describe a pipeline for structure-from-motion with mixed camera types, namely omnidirectional and perspective cameras. The steps of the pipeline can be summarized as calibration, point matching, pose estimation, triangulation and bundle adjustment. For these steps, we either propose improved methods or modify existing perspective camera methods to make the pipeline more effective and automatic when employed for hybrid camera systems.

Keywords—omnidirectional camera, hybrid camera system, feature matching, structure-from-motion

I. INTRODUCTION

Omnidirectional cameras provide 360° horizontal field of view in a single image. Major drawback of these images is that they have lower resolution than perspective images. Using perspective cameras together with omnidirectional ones could improve the resolution while preserving the enlarged view advantage. While working with such hybrid camera systems, we need to modify the approaches that are employed in systems using one type of camera.

A. Previous Work

Chen *et al.*[1] worked on calibration of a perspective-catadioptric camera system using pre-measured 3D points in the scene. Puig *et al.*[2] worked on feature point matching and fundamental matrix estimation between perspective and catadioptric camera images. For point matching, they first applied a catadioptric-to-panoramic conversion and then employed SIFT [3] between panoramic and perspective views. To eliminate the false matches they employed RANSAC [4] based on satisfying the epipolar constraint.

Ramalingam *et al.*[5] conducted a study on hybrid structure-from-motion (SfM). They used manually selected point correspondences to estimate epipolar geometry and employed midpoint triangulation method to estimate 3D point coordinates. They also compared two different bundle adjustment approaches, one minimizing the distances between projection rays and 3D points and the other minimizing the reprojection error.

B. Our Contribution

The SfM pipeline we describe is commonly used for perspective camera systems. Our contribution is bringing

together methods for the steps of this pipeline to make it effective when used for hybrid camera systems. Several of these methods were developed by us to increase the performance of the tasks in mixed camera environments.

We employ the sphere camera model [6] which is able to cover single viewpoint catadioptric systems and fisheye cameras [7] as well as perspective cameras. We calibrate our cameras according to this model by a recent calibration technique [8]. For feature point matching, we employ an algorithm which significantly increases SIFT matching performance for hybrid image pairs [9]. We robustly estimate the hybrid epipolar geometry using RANSAC and evaluate the alternatives of pose estimation methods. We propose a weighting strategy for iterative linear triangulation which improves the structure estimation accuracy. Finally, we apply sparse bundle adjustment method [10] for mixed camera types.

II. THE SfM PIPELINE

A. Camera Model and Calibration

To calibrate the cameras according to the sphere model [6], we recently proposed a calibration technique [8] in which initial intrinsic parameters are estimated linearly making use of *lifted coordinates* and estimating a 6x10 projection matrix. This method has the advantage of linear and automatic parameter initialization compared to the method of Mei and Rives [11] which also calibrates the sphere model.

B. Feature Matching

To extract the geometry between camera views, we need to match points between those views. SIFT is a popular feature detection and matching method [3]. It detects features in the so-called *scale space* comprising levels and octaves which are obtained by low-pass filtering and downsampling the original image systematically. Although this enables the detection of features at different scales, we observed that, in our hybrid case, most of the false matches in SIFT output are due to matching a high-resolution feature in perspective image to a feature in omnidirectional image which does not contain such high-resolution.

To match features in hybrid image pairs automatically, we proposed an algorithm to obtain a better SIFT feature match

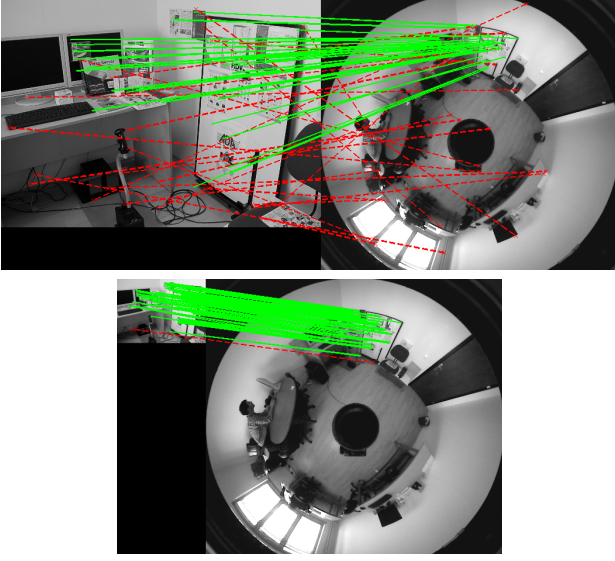


Figure 1. Matching results for an example hybrid image pair with (top) and without (bottom) the proposed preprocessing approach. Red dashed lines show the false matches, green solid lines show the correct ones.

set. Our method includes preprocessing (low-pass filtering and downsampling) perspective images before matching. In this way, many candidates from the first octaves of the perspective image are eliminated and matching produces a significantly higher true/false ratio. An example is given in Fig. 1 after applying the proposed method where true/total ratio is 54/55. For the same hybrid image pair, direct SIFT matching results in a true/total ratio of 35/60. This method also improves performance of matching between perspective images where an approximate scale ratio exists in between. The details of the algorithm and experimental results are presented in [9].

C. Robust Epipolar Geometry Estimation

Epipolar geometry between hybrid camera views was explained by Sturm [12] for mixtures of paracatadioptric (catadioptric camera with a parabolic mirror) and perspective cameras. The framework can also be extended to cameras with lens distortion due to the similarities between the paracatadioptric and division models [13]. According to these studies, a 3×4 fundamental matrix describes the relationship between a perspective and a para-catadioptric image:

$$\mathbf{q}_p^T \mathbf{F}_{pc} \hat{\mathbf{q}}_c = 0 \quad (1)$$

where \mathbf{q}_p and $\hat{\mathbf{q}}_c$ are image point coordinates in perspective and catadioptric images respectively, and $\hat{\mathbf{q}}_c$ is represented in *lifted coordinates* to linearize the equations: $\hat{\mathbf{q}}_c = (x^2 + y^2, x, y, 1)^T$.

1) *Normalization*: Normalization of image point coordinates is crucial for fundamental matrix estimation of perspective cameras [14]. We succeeded to define 4×4 T matrices for normalization of lifted coordinates ($\hat{\mathbf{q}}_{norm} = \mathbf{T}\hat{\mathbf{q}}$), so that the normalized coordinates still suit to the lifted form.

Let n be the value used for scale normalization and (c_x, c_y) be the centroid of the points in the image, lifting a normalized point leads to the 4-vector:

$$\hat{\mathbf{q}}_{norm} = \left(\frac{(x-c_x)^2}{n^2} + \frac{(y-c_y)^2}{n^2}, \frac{x-c_x}{n}, \frac{y-c_y}{n}, 1 \right) \quad (2)$$

The transformation T defined as in Eq. 3 yields $\hat{\mathbf{q}}_{norm}$ when multiplied with unnormalized lifted coordinates ($\hat{\mathbf{q}}$). Denormalization of \mathbf{F}_{pc} can also be performed linearly using T.

$$\mathbf{T}\hat{\mathbf{q}} = \begin{pmatrix} \frac{1}{n^2} & \frac{-2c_x}{n^2} & \frac{-2c_y}{n^2} & \frac{c_x^2 + c_y^2}{n^2} \\ 0 & \frac{1}{n} & 0 & \frac{-c_x}{n} \\ 0 & 0 & \frac{1}{n} & \frac{-c_y}{n} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^2 + y^2 \\ x \\ y \\ 1 \end{pmatrix} \quad (3)$$

For scale normalization in perspective images, it is suggested to normalize the point coordinates so that the RMS distance of the points from the origin is equal to $\sqrt{2}$ since it indicates the case that (x, y) coordinates are normalized to (1,1) which minimizes the difference between bare, multiplied and powered coordinate values ($1=1 \times 1=1^2$), a crucial condition for linear estimation of fundamental matrix. We investigated the similar optimal case for hybrid pairs and conducted experiments with hybrid pairs of real and simulated images [15]. Best results were obtained again when RMS distance is normalized to $\sqrt{2}$.

2) *Outlier elimination*: After initial detection of matches, RANSAC [4] algorithm based on this hybrid epipolar relation is used to eliminate false matches. Since \mathbf{F}_{pc} has 12 elements, the minimum number of correspondences needed to estimate \mathbf{F}_{pc} is $12-1(\text{scale factor})=11$. As in perspective camera case, we define a distance (d) to distinguish outliers and inliers, where the points closer to its corresponding epipolar line/curve than d are called inliers. In our experiments, we use $d = d_l + d_c$, where d_l is the point-to-line distance in the perspective image and d_c is the point-to-conic distance in the catadioptric image.

3) *Pose estimation*: The rotation and translation between camera views are extracted from the essential matrix (E) with the technique given in [14]. We analyzed two methods for the estimation of E. First option is directly estimating E with the calibrated 3D rays of the correspondences in the RANSAC output. The other option is estimating \mathbf{F}_{pc} with RANSAC and then extracting E from \mathbf{F}_{pc} using the relations given in [13].

In our experiments of comparing these two options, directly estimating E resulted in less 3D and reprojection error [15]. Extracting E from \mathbf{F}_{pc} is more vulnerable to noise. This is not surprising since lifted coordinates are used to compute \mathbf{F}_{pc} which increases the impact of noise and \mathbf{F}_{pc} has 12 elements whereas E has 9. Another reason for not

choosing E from F_{pc} approach is that it is only possible for para-catadioptric systems.

D. Triangulation

Triangulation is the step of estimating 3D coordinates for the matched 2D points using camera poses. We improved *iterative linear-Eigen* triangulation method for effective use in hybrid SfM. According to the comprehensive study by Hartley and Sturm [16], *iterative linear-Eigen* is one of best triangulation methods for Euclidean reconstruction. It is superior to midpoint method and non-iterative linear methods. For projective reconstruction, polynomial triangulation method performs better, however it requires a considerable amount of computation time and not easily generalizable to more than two images.

The perspective cameras in mixed systems tend to have higher resolution than the omnidirectional ones. To benefit from their resolution, we increased the weight of measurements (rows in linear triangulation) coming from perspective images. We employ this triangulation method with calibrated 3D rays instead of raw pixels. Since the projection in omnidirectional cameras can not be expressed linearly as in perspective cameras, hybrid triangulation can be performed with the 3D rays outgoing from the effective viewpoints of the cameras. Please note that, for two perspective cameras, triangulation can be performed on pixels (not on 3D rays) and since the reprojection error in the image is minimized, the iterative linear triangulation supports the zoomed image without requiring an extra weighting for the zoomed images.

With the mentioned weighting strategy, we observed improvement in the accuracy of estimated 3D coordinates. The details and experiment results can be found in [15]. To give an idea about the amount of improvement, in a simple case where measurements from the perspective camera are multiplied by the ratio of the scales of the objects in perspective and omnidirectional images, the 3D location error decreased by 3.2-9.6% when scale ratio is 2-4.

E. Adding a New View

To perform multi-view SfM, we employed the approach proposed by Beardsley *et al.*[17]. In this approach, when a sequence of views is available, initially SfM is applied for the first two views. Then, for each new view, features are detected and matched with the previous view, which are associated with already reconstructed 3D points. The projection matrix of the new view is computed using these final 2D-3D matches.

An alternative approach being used for multi-view SfM is *projective factorization* and it constructs a measurement matrix which contains the projections of all 3D points in all cameras [18], [19]. The result of this method is always a projective reconstruction and requires conversion to an Euclidean reconstruction using additional methods. Most importantly, the projection of omnidirectional images

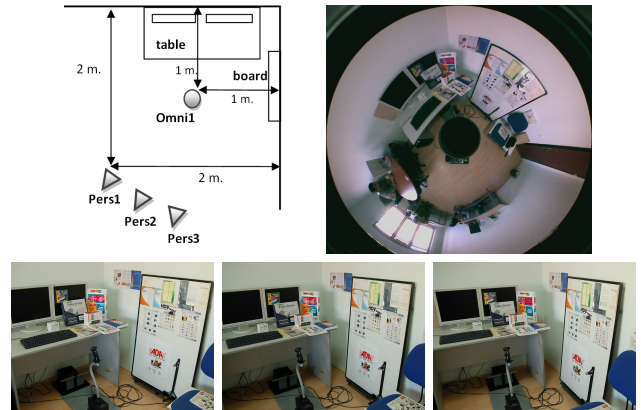


Figure 2. Images of the hybrid multi-view SfM experiment. Omni1 is seen on top-right, Pers1, Pers2 and Pers3 are at bottom from left to right. Sketch on top-left shows the locations and orientations of cameras w.r.t. the scene.

can not be written in a linear fashion together with the perspective images. Thus, this multi-view approach is not suitable for our case.

F. Bundle Adjustment

Sparse bundle adjustment (SBA) method proposed by Lourakis and Argyros [10] has become popular in the community due to its capability of solving huge minimization problems (with many cameras and 3D points) in a reasonable time. We employed this method for our system of mixed cameras. We modified the projection function with the sphere model projection and intrinsic parameters with sphere model parameters.

III. MULTI-VIEW SfM EXPERIMENT

We present here a hybrid multi-view SfM experiment in which the entire proposed pipeline is applied. Initial structure estimation was performed with Pers1-Pers2 pair, then Pers3 and Omni1 views were added (Fig. 2). 75 feature points are common in all images. Estimated coordinates of these points and estimated camera positions are shown in Fig. 3, where O_o shows the center of omnidirectional camera and O_i shows the center of i^{th} perspective camera. Please compare the estimated camera positions and orientations (Fig. 3a) with the actual ones (Fig. 2, top-left).

We performed SBA on this structure (scene point coordinates) and camera parameters. The reprojection errors before and after SBA (in pixels) are given in Table I. We infer from the table that the reprojection errors are considerably decreased after SBA. The error before SBA for the omnidirectional image is higher than the perspective images due to the fact that the number of common points between omnidirectional and perspective images is less when compared to the number of common points between two perspective images, which decreases the accuracy of pose estimation.

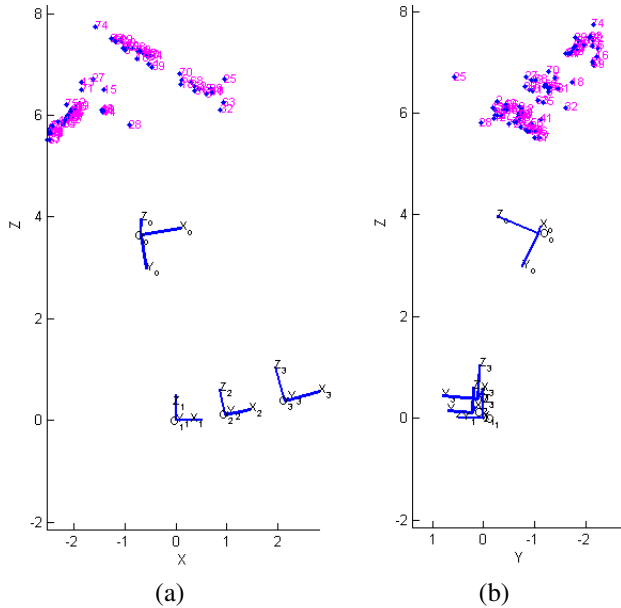


Figure 3. Estimated camera positions, orientations and scene points for the hybrid multi-view SfM experiment. (a) top-view (b) side-view.

Table I
THE REPROJECTION ERROR MEAN VALUES BEFORE AND AFTER SBA
(IN PIXELS).

	Pers1	Pers2	Pers3	Omni1
Before SBA	0.49	0.48	0.53	0.97
After SBA	0.28	0.23	0.29	0.39

IV. CONCLUSIONS

We demonstrated that, with the described pipeline and proposed methods, SfM with hybrid camera systems can be performed as easily as in perspective camera systems. We employed sphere camera model which enabled us not to change camera model between camera types. We also employed the calibration technique that we developed for this camera model. For feature matching, we proposed an improved method which is based on SIFT. With this method, automatic hybrid matching becomes possible. For triangulation step as well we propose an improvement to increase accuracy.

For other steps, namely, coordinate normalization, RANSAC for hybrid epipolar geometry estimation, multi-view SfM and sparse bundle adjustment, we tested and modified existing approaches to be used for hybrid systems.

REFERENCES

[1] X. Chen, J. Yang, and A. Waibel, "Calibration of a hybrid camera network," in *Proc. of IEEE International Conference on Computer Vision (ICCV)*, 2003.

[2] L. Puig, J. Guerrero, and P. Sturm, "Matching of omnidirectional and perspective images using the hybrid fundamental matrix," in *Workshop on Omnidirectional Vision*, 2008.

[3] D. Lowe, "Distinctive image features from scale invariant keypoints," *Int. J. of Computer Vision*, vol. 60, 2004.

[4] M. Fischler and R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24(6), 1981.

[5] S. Ramalingam, S. Lodha, and P. Sturm, "A generic structure-from-motion algorithm for cross-camera scenarios," in *Workshop on Omnidirectional Vision*, 2004.

[6] C. Geyer and K. Daniilidis, "A unifying theory for central panoramic systems," in *European Conference on Computer Vision (ECCV)*, 2000.

[7] X. Ying and Z. Hu, "Can we consider central catadioptric cameras and fisheye cameras within a unified imaging model?" in *European Conference on Computer Vision (ECCV)*, 2004, pp. 442–455.

[8] Y. Bastanlar, L. Puig, P. Sturm, J. Guerrero, and J. Barreto, "Dlt-like calibration of central catadioptric cameras," in *Proc. of Workshop on Omnidirectional Vision*, 2008.

[9] Y. Bastanlar, A. Temizel, and Y. Yardimci, "Improved sift matching for image pairs with scale difference," *Electronics Letters*, vol. 46(5), pp. 346–348, 2010.

[10] M. Lourakis and A. Argyros, *The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package based on the LM Algorithm*. FORTH-ICS Technical Report, TR-340, 2004.

[11] C. Mei and P. Rives, "Single viewpoint omnidirectional camera calibration from planar grids," in *International Conference on Pattern Recognition (ICPR)*, 2007.

[12] P. Sturm, "Mixing catadioptric and perspective cameras," in *Workshop on Omnidirectional Vision*, 2002.

[13] J. Barreto and K. Daniilidis, "Epipolar geometry of central projection systems using veronese maps," in *Proc. of Computer Vision and Pattern Recognition*, 2006, pp. 1258–1265.

[14] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge Univ. Press, 2004.

[15] Y. Bastanlar, *Structure-from-Motion for Systems with Perspective and Omnidirectional Cameras*. Ph.D. Thesis, Middle East Technical University, 2009.

[16] R. Hartley and P. Sturm, "Triangulation," *Computer Vision and Image Understanding*, vol. 68(2), pp. 146–157, 1997.

[17] P. Beardsley, A. Zisserman, and D. Murray, "Sequential updating of projective and affine structure from motion," *International Journal of Computer Vision*, vol. 23(3), pp. 235–259, 1997.

[18] P. Sturm and B. Triggs, "A factorization based algorithm for multi-image projective structure and motion," in *Proc. of European Conference on Computer Vision (ECCV)*, 1996, pp. 709–720.

[19] D. Martinec and T. Pajdla, "Structure from many perspective images with occlusions," in *Proc. of European Conference on Computer Vision (ECCV)*, vol. 2, 2002, pp. 355–369.