Multi-view Structure-from-Motion for Hybrid Camera Scenarios

Y. Bastanlar[a,1,*], A. Temizel[a], Y. Yardimci[a], P. Sturm[b]

[a] *Informatics Institute, Middle East Technical University, 06531, Ankara, Turkey*
[b] *INRIA Rhone-Alpes and Laboratoire Jean Kuntzmann, Grenoble, France*

Abstract

We describe a pipeline for structure-from-motion (SfM) with mixed camera types, namely omnidirectional and perspective cameras. For the steps of this pipeline, we propose new approaches or adapt the existing perspective camera methods to make the pipeline effective and automatic. We model our cameras of different types with the sphere camera model. To match feature points, we describe a preprocessing algorithm which significantly increases scale invariant feature transform (SIFT) matching performance for hybrid image pairs. With this approach, automatic point matching between omnidirectional and perspective images is achieved. We robustly estimate the hybrid fundamental matrix with the obtained point correspondences. We introduce the normalization matrices for lifted coordinates so that normalization and denormalization can be performed linearly for omnidirectional images. We evaluate the alternatives of estimating camera poses in hybrid pairs. A weighting strategy is proposed for iterative linear triangulation which improves the structure estimation accuracy. Following the addition of multiple perspective and omnidirectional images to the structure, we perform sparse bundle adjustment on the estimated structure by adapting it to use the sphere camera model. Demonstrations of the end-to-end multi-view SfM pipeline with the real images of mixed camera types are presented.

Keywords: Omnidirectional cameras, Hybrid camera systems, Feature matching,

*Corresponding author
Email addresses: yalinbastanlar@iyte.edu.tr (Y. Bastanlar), atemizel@ii.metu.edu.tr (A. Temizel), yardimy@ii.metu.edu.tr (Y. Yardimci), peter.sturm@inrialpes.fr (P. Sturm)
[1] Presently, the author is with Computer Eng. Dept., Izmir Institute of Technology, Izmir, Turkey.

Epipolar geometry, Multi-view, Structure-from-motion

1. Introduction

Omnidirectional cameras provide a 360° horizontal field of view in a single image, which is an important advantage in many application areas such as surveillance [1, 2], robot navigation [3, 4] and 3D reconstruction [5, 6]. Point correspondences from a variety of angles provide more stable structure estimation [7] and degenerate cases like viewing only a planar surface are less likely to occur. Omnidirectional images can provide omni-present correspondences when the fields of view of perspective images do not overlap as we will discuss in this paper. A major drawback of these cameras is their lower spatial resolution than perspective cameras due to their large field of view. Using perspective cameras together with the omnidirectional ones could improve the resolution while preserving the advantage of an enlarged field of view. A possible scenario is 3D reconstruction in which omnidirectional cameras provide low resolution background reconstruction whereas the images of perspective cameras are used for modeling specific objects in the foreground. Another possible application opportunity for hybrid SfM is hybrid surveillance systems, for instance using a pan-tilt-zoom camera with an omnidirectional camera [1]. Enhancement of these systems by 3D structure and location estimation algorithms is possible without increasing the number of cameras.

For 3D reconstruction with such hybrid camera systems, we need to adapt the steps that are employed in systems using a single type of camera. In Fig. 1, an SfM pipeline which is commonly used for perspective camera systems is given. We investigate the applicability of this pipeline to hybrid camera systems and we propose improved or modified methods for different steps of this pipeline when needed.

Regarding the previous studies on hybrid systems, Adorni *et al.*[3] used a hybrid system for the obstacle detection problem in robot navigation. Chen and Yang [8] developed a region matching algorithm for hybrid views based on planar homographies. The epipolar geometry between hybrid camera views was first explained by Sturm [9] for mixtures of paracatadioptric (catadioptric camera with a parabolic mir-

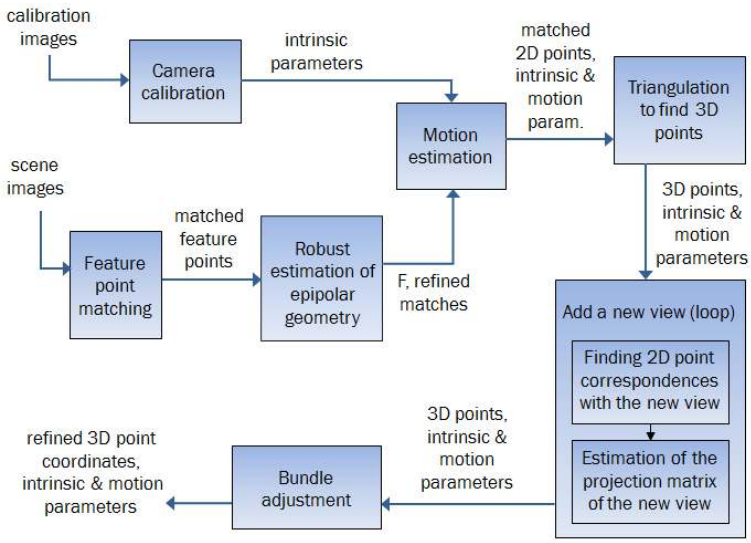


Figure 1: Steps of the applied SfM pipeline.

ror) and perspective cameras. The framework was extended to catadioptric cameras with hyperbolic mirrors and cameras with lens distortion by Barreto and Daniilidis [10]. Puig *et al.*[11] 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 employed regular SIFT [12] between panoramic and perspective views. They employed RANSAC [13] based on satisfying the epipolar constraint and compared the representation capabilities of 3x4, 3x6 and 6x6 hybrid fundamental matrices for mirrors with varying parameters.

To our knowledge, the only work on hybrid SfM was conducted by Ramalingam *et al.*[14]. They employed a highly generic non-parametric imaging model where the cameras are modeled with sets of projection rays. They mentioned that directly applying SIFT [12] did not provide good results for their fisheye-perspective image pairs and used manually selected feature point correspondences to estimate the epipolar geometry. They employed the midpoint method for triangulation to estimate 3D point coordinates. They also tested two different bundle adjustment approaches, one minimizing the distances between projection rays and 3D points and the other minimizing reprojection error, concluding that both approaches are comparable to each other.

In our work, we employ the sphere camera model [15] which covers central (single-viewpoint) catadioptric systems as well as perspective cameras. The details are presented in Section 2. The SfM pipeline described here applies to all the cameras that can be modeled with the sphere model. The proposed methods for point matching and triangulation can be used with the cameras beyond the scope of the sphere camera model, since they do not employ this camera model.

Widely accepted feature matching methods (eg. SIFT [12], MSER [16]) do not perform well when they are directly employed for hybrid camera images [11, 14]. Main reasons are the resolution difference and the distortion of features between the images of different camera types. Our analysis showed that, most of the false matches in SIFT output are due to matching a high-resolution feature in the perspective image to a feature in the omnidirectional image which does not have such high-resolution. We propose an algorithm that preprocesses the perspective images before matching. In this way, the probability of matching the features between the incorrect scales (octaves) decreases and SIFT matching produces a significantly higher true-positive ratio allowing us to perform automatic omnidirectional-perspective matching. We performed tests on a total of 20 image pairs taken from different scenes and with different omnidirectional (both catadioptric and fisheye) cameras. To decrease the effect of distortion in hybrid image pairs, we evaluate the use of virtual camera plane (VCP) images and include VCP-perspective matching to our experiments. Experimental results, given in Section 3, indicate the success of our method.

We employ RANSAC [13] to robustly compute the hybrid fundamental matrix (F) which requires the usage of *lifted coordinates* for linear estimation [9, 10]. We introduce normalization matrices for lifted coordinates so that normalization and denormalization can be performed linearly. We compare two options for motion estimation, one is directly estimating the essential matrix (E) with the calibrated 3D rays, the other option is estimating the hybrid F and then extracting E from F. We give details of our analysis in Section 4.

The only previous study including hybrid camera triangulation is using the midpoint method [14], however it was shown that iterative linear methods are superior to the midpoint method [17]. We propose a weighting strategy for the *iterative linear-*

Eigen triangulation method to improve its 3D location estimation accuracy by trusting the (high resolution) perspective image more when employed for hybrid image pairs (Section 5).

In Section 6, we describe how we perform multi-view SfM. Briefly, we employ the approach of adding views to the structure [18] and to refine the final 3D point coordinates and camera motion parameters, we adapt the *sparse bundle adjustment* method [19] by modifying its projection function to be used with the sphere camera model.

We present the results of our experiments for the individual steps of the SfM pipeline within the related sections. In Section 7, we present the demonstrations of the complete pipeline, i.e. multi-view hybrid SfM with real images, to show that our approach is working effectively in real world scenarios. We present two scenarios, also mentioned at the beginning of this section, where employing a hybrid camera system is advantageous. One of them is a surveillance setup where the scene can be dynamic and images are captured simultaneously. Thus, a mobile camera can not be used and it is not practical to use many perspective cameras to cover the whole scene. Section 7.1 presents such a scenario in which an omnidirectional camera is used in conjunction with a limited number of perspective cameras that do not view the same part of the scene. Such hybrid systems are becoming more widespread with the increased demand for video surveillance. We demonstrate how an omnidirectional camera can combine the 3D structures viewed by two or more perspective cameras with no overlapping views. Section 7.2 presents a second scenario in which two omnidirectional images are used to provide low resolution background reconstruction whereas several perspective views are used for modeling the objects at the foreground. The hybrid method alleviates the need for a mobile camera or a network of cameras for background reconstruction. In Section 7.3, we draw the reader's attention to another advantage of the hybrid systems and we demonstrate that adding omnidirectional cameras to perspective SfM scenarios increases the accuracy of motion estimation. Finally, in Section 7.4, we present an outdoor experiment, in which an image sequence from a captured video was used, to investigate the applicability of our hybrid SfM in other realistic scenarios.

The work presented here mainly comprises the research included in the first au-

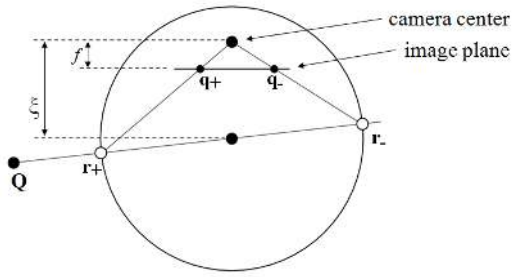


Figure 2: Projection of a 3D point to two image points in the sphere camera model.

thor's dissertation [20] and some of the experimental results are presented in [21].

2. Camera Model and Calibration

We use the sphere camera model by Geyer and Daniilidis [15] which was introduced to model central catadioptric cameras. Later, this model was extended to cover perspective cameras with lens distortions [22]. The model, comprises a unit sphere and a perspective camera and the projection of 3D points can be performed in two steps (Fig. 2). The first one is the projection of point $\mathbf{Q}$ in 3D space onto a unitary sphere and the second one is the projection from the sphere to the image plane. The first projection gives rise to two intersection points on the sphere, $\mathbf{r}_{\pm}$. The one that is visible to us is $\mathbf{r}_{+}$ and its projection on the image plane is $\mathbf{q}_{+}$. This second projection is realized by $\mathbf{q}_{+} \sim \mathsf{K}\mathbf{r}_{+}$ where K is a projection matrix including the intrinsic and extrinsic parameters of the perspective camera embedded in the sphere model. The sphere model covers all central catadioptric cameras described by the distance between the camera center and the center of the sphere, ξ. $\xi = 0$ for perspective, $\xi = 1$ for para-catadioptric, $0 < \xi < 1$ for hyper-catadioptric cameras. This projection geometry is described in detail in [23].

Several methods were proposed for the calibration of catadioptric systems. Some of them consider estimating the parameters of the mirror together with the camera parameters [24, 25, 26, 27], some others calibrate outgoing rays based on a radial distortion model [28, 29, 30]. Since we use the sphere camera model for our projections, we calibrate our cameras with this model. Mei and Rives [31] developed a MATLAB cal-

ibration toolbox for the sphere model. It requires user input for initialization of focal length and principal point. The user also needs to define the type of the mirror and the toolbox is not able to calibrate perspective cameras.

A recent contribution on the sphere model calibration is given in [32], authors of which contributed to this paper as well. In this technique, initial intrinsic parameters are estimated linearly making use of *lifted coordinates* and estimating a 6x10 projection matrix. Then the parameters are optimized to reach the minimum reprojection error. This algorithm requires a 3D calibration pattern and brings the advantage of linear and automatic parameter initialization.

3. Feature Matching

To match the features in hybrid image pairs automatically, we describe a preprocessing algorithm to be applied with SIFT [12] matching. Matching performance decreases with distortion of features due to increasing baseline length or changing camera geometry. SIFT detects features at different scales and matches them regardless of their scales. We also observed low matching accuracy when there is a major scale difference between the two images. These conditions especially apply to our hybrid image pairs where there is an inherent resolution difference and distortion between the images of different camera types. Most of the false matches in the SIFT output are due to matching a high-resolution feature in the perspective image to a feature in the omnidirectional image. Let us explain this phenomenon by an example. Table 1 shows the number of extracted features in the so-called *octaves* of an hybrid image pair. There is an approximate ratio between the scales of true correspondences (SR=$\sigma_{pers}/\sigma_{omni}$), which is indicated in the table with arrows. SIFT extracts many features (nearly 3000) at the first two (high-resolution) octaves. Due to the distortion in the images, some of the excessive number of candidates from the first two octaves of the perspective image are incorrectly selected as the best match of the features in the omnidirectional image since they accidentally have close enough feature orientation vectors. If there were no candidates from incorrect scales, these false matches would have been prevented. For the given image pair, there are 25 false matches out of 60 and 23 of these false matches

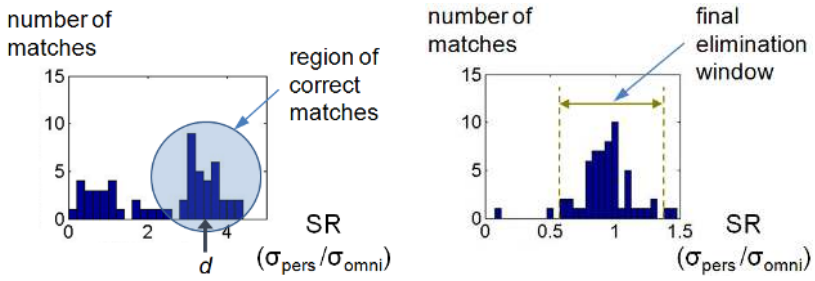
Octave	Approximate scale(σ) in SIFT scale space	Omnidirectional image (1024x960)	Perspective image (1100x800)	Perspective, preprocessed
-1	1	1365	2459	288
0	2	489	463	97
1	4	202	174	23
2	8	68	76	5
3	16	23	20	0
4	32	4	5	-

Table 1: Number of SIFT features detected in different octaves of a perspective-omnidirectional image pair. Corresponding octaves of correct matches are indicated with arrows. Last column shows the number of features when the perspective image is downsampled by 3.6 (both in horizontal and vertical axis) following a low-pass filtering operation.

have an SR$\leq$2.0, whereas the average SR of true matches is 3.57.

The histogram of the example hybrid image pair in Table 1 is shown in Fig. 3a. The accumulation on the left (matches with SR$<$2.0) is explained by the false matches due to matching features in the first two octaves of the perspective image. False matches can be eliminated by defining a window around the peak and rejecting the matches outside. A similar elimination was performed in [33, 34] where only perspective image pairs were considered. Yi *et al.*[33] worked on perspective images with approximately the same scale and eliminated the matches with scale differences outside the selected window. In the method proposed by Alhwarin *et al.*[34], the octave pair which yields the maximum number of matches is detected and the matches from other octave pairs are rejected.

However, instead of performing such an elimination after matching, we suggest improving the SIFT matching output by processing the high resolution image so that its resolution matches that of the lower resolution image. Doing so not only eliminates the false matches but also increases the number of correct matches. This can be achieved by low-pass filtering and downsampling the perspective image in a hybrid pair. With this preprocessing, the scale ratio of matching features becomes close to 1 and the possibility of matching valuable features in the omnidirectional image with the features in the correct octaves of the perspective image considerably increases since the high-resolution candidate features in the perspective image are already eliminated.



(a) (b)

Figure 3: Histogram of SR for the matches in the perspective-omnidirectional image pair given in Table 1. (a) SIFT applied on original image pair, (b) SIFT applied on the preprocessed perspective image and the original catadioptric image.

We selected the downsampling factor from the histogram as the mean of the most dominant Gaussian in the mixture (Fig. 3a), because the SIFT scale space ratio also reveals the scale ratio of features in the images. To avoid aliasing, we low-pass filter the perspective image before downsampling. We selected the cut-off frequency as $2.5/\sigma$ in the frequency domain and the standard deviation of the Gaussian low-pass filter becomes $\sigma = 2.5d/\pi$ where d is the downsampling factor. Fig. 3b shows the SR histogram when the perspective image is low-pass filtered and downsampled as explained. This matching resulted in a true/total ratio of 56/60.

We can further restrict the scale to remove a few false matches with improper SR. To do this we define a window around the mean scale ratio $\overline{\mathrm{SR}}$ and discard the matches outside the window. For our algorithm we chose the bounds of the restriction window as [$0.6\overline{\mathrm{SR}}$,$1.4\overline{\mathrm{SR}}$] (Fig. 3b). After this final elimination, true/total ratio becomes 54/55 for the given example. Visual matching results for the example image pair before and after the proposed method are given in Fig. 4.

Let us summarize the steps of the proposed matching method:

1. Following an initial matching, extract the downsampling factor (d) from the histogram of scale ratios (eg. Fig. 3a).
2. Low-pass filter the perspective image with a Gaussian filter having a $\sigma = 2.5d/\pi$ and downsample it by d.

3. Apply SIFT matching between the preprocessed perspective image and the original omnidirectional image.
4. Perform final elimination on the histogram of scale ratios (SR) of the final matching (eg. Fig. 3b).

If we directly apply scale restriction (Step 4) without preprocessing, similar to the approaches in [33, 34], the resultant true/total ratio is 32/34. We will refer to this approach as *Scale Restricted SIFT*. Although most of the false matches are eliminated with this scale restriction, our method keeps a higher number of correct matches with a true/total of 54/55. In our method, many high-resolution candidate features are already eliminated before the final matching and they no longer act as attraction centers. A performance comparison between our method and *Scale Restricted SIFT* is presented with experimental results in Section 3.2. It is important to keep as many true matches as possible in most computer vision applications, especially for structure-from-motion.

An earlier version of our method concentrated on the feature matching for perspective image pairs [35]. We showed that the proposed SR detection and preprocessing approach works on perspective image pairs with wide-baseline distortion.

In this paper, we describe the reasons of false matches in SIFT on hybrid pairs. We employ virtual camera plane (VCP) images to decrease the effect of distortion between hybrid images. The elimination window in the final elimination (Step 4) can be made adaptive. Due to the decreasing resolution towards the center of a catadioptric image, the scale of the objects close to the center is smaller than the scale of those at the periphery of the image. In order to accommodate varying scale within the image, the upper and lower boundaries of the elimination window are decreased (expected SR is lower) for a point that is closer to the periphery. If a point is closer to the center, the boundaries are increased. This adjustment resulted in a marginal improvement and is suggested when the feature points are scattered from the center to the periphery of the catadioptric image. The integration of the point matching step to the SfM pipeline is presented in Section 4.2 where the remaining false matches that do not conform to the hybrid epipolar constraint are eliminated.

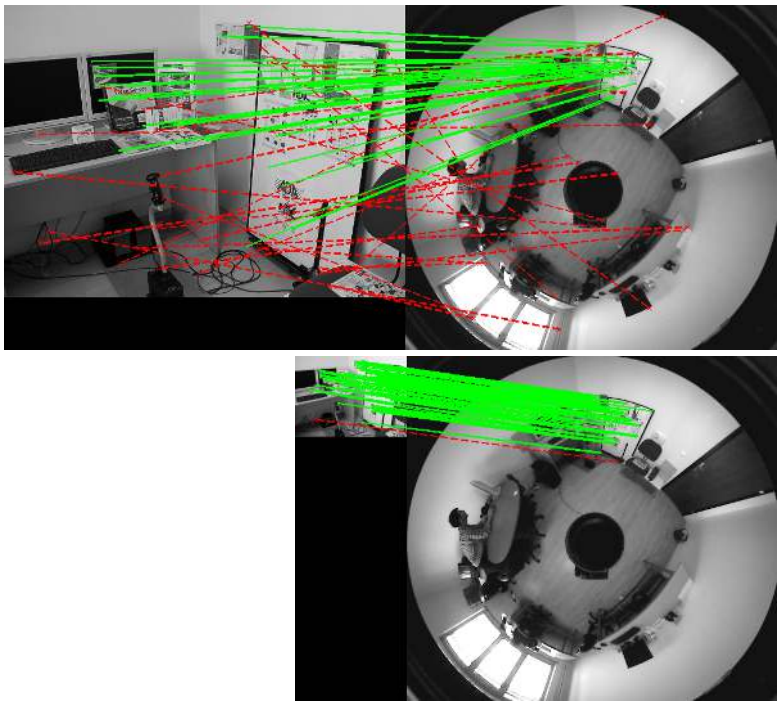


Figure 4: SIFT matching result for the sample hybrid image pair before (top) and after (bottom) the proposed approach. Final true/total ratio is 54/55, whereas it was 35/60 as the output of regular SIFT. Red dashed lines show the false matches, green solid lines show the correct ones.

3.1. Creating VCP Images

In addition to the omnidirectional-perspective matching, we investigate matching virtual camera plane (VCP) images with perspective images (Fig. 5). The motivation here is to decrease the effect of the distortion of the features between the different camera types. A VCP image is produced by unwarping a certain region of the omnidirectional image to generate a perspective image (an example exists in [36]). Many such VCP images can be generated. To perform matching with a certain perspective image, we generate one VCP image with a certain viewing direction (azimuth), a vertical angle and a distance from the viewpoint (zoom) so that the VCP field of view is as close as possible to that of the perspective image.

Since we created VCP images to match the size of the perspective images, the SR of the true matches is already close to 1 and no downsampling is needed. However,

Figure 5: Representation of VCP matching procedure.

we still need to low-pass filter the perspective image to prevent SIFT from extracting high-resolution spurious attractors.

3.2. Experiments

To test the robustness of the proposed algorithm, we performed experiments on a total of 20 image pairs taken in different scenes with three different omnidirectional cameras. We used the SIFT implementation of Andrea Vedaldi [2] with a modification to yield one-to-one matching.

Fig. 6 shows the images used in one of the experiments. Omnidirectional images are captured with Remote Reality S80 Optic [3]. Matching results are given in Table 2, where we can compare the three approaches: direct omnidirectional-perspective matching, matching after downsampling the perspective image and VCP-perspective matching. We plot the ratio of true/total matches in Fig. 7 to provide easiness in comparison. To keep the number of matched points same for different trials of an image pair, we adjusted the matching threshold of SIFT, which defines the strength of the matched point with respect to the second candidate match.

We observe that the matching performance decreases with increasing baseline (from Pers1 to Pers3) for all approaches. This is the natural performance decrease due to the wide baseline and occurs regardless of being a hybrid pair or not. It can be seen that the proposed method (both with and without the VCP approach) outperforms SIFT matching applied on original images. We also observe that the VCP approach exhibits slightly better results since it attempts to solve both the distortion and the resolution

[2] http://vision.ucla.edu/∼vedaldi/code/sift/sift.html

[3] http://www.remotereality.com

Figure 6: Images used for the experiment results of which is given in Table 2. Top row is Omni1 and VCP image generated from it (VCP1), second row is Omni2 and VCP2, bottom row is the perspective images, Pers1, Pers2 and Pers3. Omnidirectional images have a size of 1024x960 pixels, whereas perspective and VCP images are 1100x800 pixels. Please note that the hybrid image pair used in Table 1, Fig. 3 and Fig. 4 is Omni2-Pers2.

difference problems.

The last column of Table 2 shows the true/false ratios obtained after final elimination. The values in this column support the observation that the proposed method (with or without VCP) provides a higher number of correct matches, when compared to *Scale restricted SIFT* (SRS). The values in Pers*N*-Omni*M* rows refer to the SRS result whereas the values in Pers*N* σA dB - Omni*M* rows are the results of our method. Regarding the Pers2-Omni2 pair for instance, our method results in 54/1 true/false matches, which overwhelms the number of true matches in the result of SRS (32/2). At the end of this section, we compare SRS with the proposed method for all the image pairs (Table 3).

Fig. 8 shows the experimental results in graphs for another catadioptric-perspective

Image pairs	no. of matches	true/false (true/total %)	true/false final
Pers1 - Omni1	100	97/3 (97%)	84/1
Pers1 σ1.5 d1.65 - Omni1	100	99/1 (99%)	94/0
Pers1 σ1.5 - VCP1	100	99/1 (99%)	99/0
Pers2 - Omni1	75	56/19 (75%)	51/7
Pers2 σ1.5 d1.65 - Omni1	75	70/5 (93%)	69/1
Pers2 σ1.5 - VCP1	75	73/2 (97%)	73/1
Pers3 - Omni1	60	42/18 (70%)	39/9
Pers3 σ1.5 d1.65 - Omni1	60	50/10 (83%)	50/6
Pers3 σ1.5 - VCP1	60	57/3 (95%)	57/1
Pers1 - Omni2	80	63/17 (79%)	62/1
Pers1 σ2.5 d3.6 - Omni2	80	80/0 (100%)	80/0
Pers1 σ2.5 - VCP2	80	80/0 (100%)	80/0
Pers2 - Omni2	60	35/25 (58%)	32/2
Pers2 σ2.5 d3.6 - Omni2	60	56/4 (93%)	54/1
Pers2 σ2.5 - VCP2	60	56/4 (93%)	56/3
Pers3 - Omni2	45	15/30 (33%)	15/1
Pers3 σ2.5 d3.3 - Omni2	45	35/10 (78%)	32/6
Pers3 σ2.5 - VCP2	45	37/8 (82%)	35/3

Table 2: Matching results for the image pairs given in Fig.6. Pers$N$ σA dB indicates that Pers$N$ image was low-pass filtered with $\sigma = A$ Gaussian filter and downsampled by a factor of B in each direction. N in VCPN indicates the index of omnidirectional image that the VCP is generated from.

camera pair, where the catadioptric images are taken with 0-360 Panoramic Optic [4]. We again have two omnidirectional and three perspective images. When compared to the direct SIFT matching, the true/total match ratio increases for both downsampling and VCP approaches. However, this time the VCP approach is distinctively better. We relate this result to the existence of repetitive patterns in the scene such as windows and repeating structures on the facade. Distorted features in the omnidirectional image are more vulnerable to being distracted by similar-looking features. The VCP approach produces more distinctive candidates for the features in the perspective image, resulting

[4] http://www.0-360.com

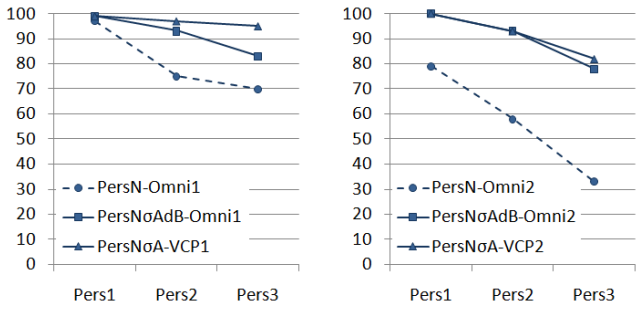


Figure 7: The true/total match ratios (in percentage) for images in Fig. 6. These graphs are also the plotted versions of the information given in Table 2. Omni1-Pers$N$ (on the left) and Omni2-Pers$N$ (on the right).

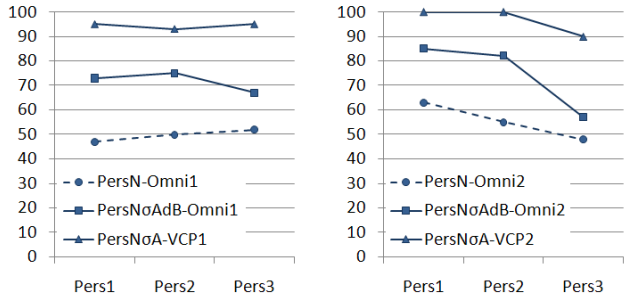


Figure 8: The true/total match ratios (in percentage) for the second omnidirectional-perspective matching experiment. Examine with Fig. 9.

in higher matching accuracy. Fig. 9 shows the correct and false matches of Pers1-Omni1 pair for the three compared matching approaches.

To investigate if the proposed approach is also valid for cameras with fisheye lenses, we conducted experiments with a Fujinon FE185C046HA 185° fisheye lens. Matching results are shown in Fig. 10 where the improvement with the proposed approaches can be easily observed. When compared to the results of experiments with catadioptric cameras, we are able to say that the performance of the VCP approach did not change but the performance of the downsampling approach increased. Fig. 11 shows the results of direct SIFT matching and downsampling approaches for Pers2-Fish1 pair.

We combine the results of 12 catadioptric-perspective and 8 fisheye-perspective image pairs in Table 3 which shows the average FP rate (# false positives / # detected matches) and TP (# true-positives) for the compared methods. In addition to the three

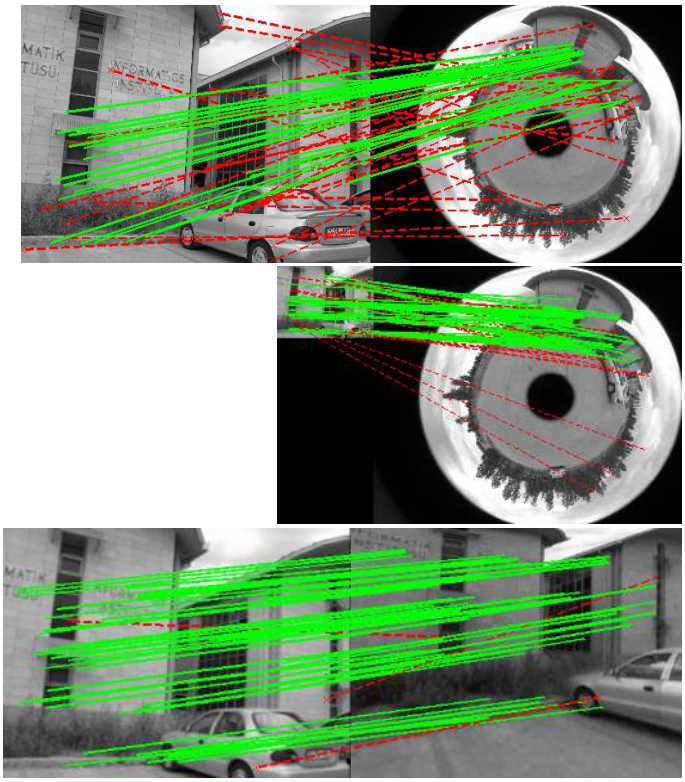


Figure 9: Matching results for the Pers1-Omni1 pair of the experiment in Fig. 8 for the three compared matching approaches: direct omnidirectional-perspective matching (top), matching after downsampling perspective image (middle) and VCP-perspective matching (bottom). Red dashed lines indicate false matches, whereas green lines indicate correct ones. From top to bottom, there are 32, 16 and 3 false matches out of 60 matches, respectively.

mentioned methods, we include *Scale restricted SIFT* in which the elimination of false matches is applied after direct SIFT matching (eg. the last column of Table 2) and can be associated with the approaches given in [33, 34]. The scale ratio between the images varies between 1.5-4.2 and the number of detected matches varies between 50 and 100.

The FP rate is very low for both *Scale restricted SIFT* and our method compared to direct SIFT matching. When the number of TP is considered, our method outperforms SIFT and *Scale restricted SIFT*. For catadioptric-perspective pairs, our method with VCP produces better results. For hybrid pairs including fisheye images, on the other hand, our method without VCP performs as well as our method with VCP conversion.

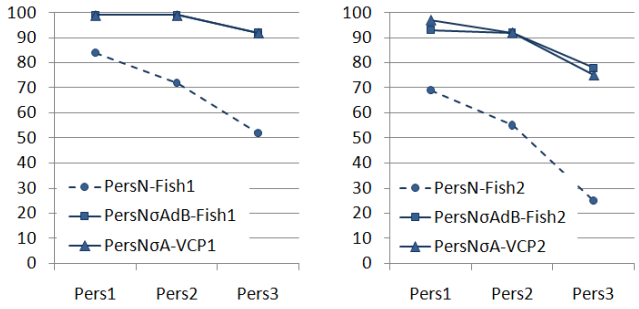


Figure 10: The true/total match ratios (in percentage) for the fisheye-perspective matching experiment. Examine with Fig. 11.

	Catadioptric		**Fisheye**	
	FP rate	TP	FP rate	TP
SIFT	0.39	42.1	0.30	56.2
Scale restricted SIFT	0.11	38.7	0.06	54.6
Our method	0.12	52.8	0.02	74.0
Our method + VCP	0.03	61.9	0.02	74.9

Table 3: Matching results including all hybrid pairs. Represented separately for pairs including catadioptric cameras and fisheye cameras.

We infer from the table that the proposed approach increases the number of correct matches rather than just eliminating the false matches afterwards. The number of correct matches is important for healthy fundamental matrix (F) estimation. The elimination of a few remaining false matches that do not conform to the epipolar geometry is also performed during the estimation of F with RANSAC. The reader will observe that RANSAC on the *Scale restricted SIFT* match set does not completely eliminate false matches and retains fewer correct matches (Section 4.2).

The proposed algorithm of preprocessing the perspective images method works with different hybrid camera types as shown by experiments. Thus, we can say that the proposed technique of extracting parameters of low-pass filtering and downsampling is versatile for omnidirectional cameras to a large extent. The reader can reach an analysis on the sensitivity of preprocessing parameters in [20].

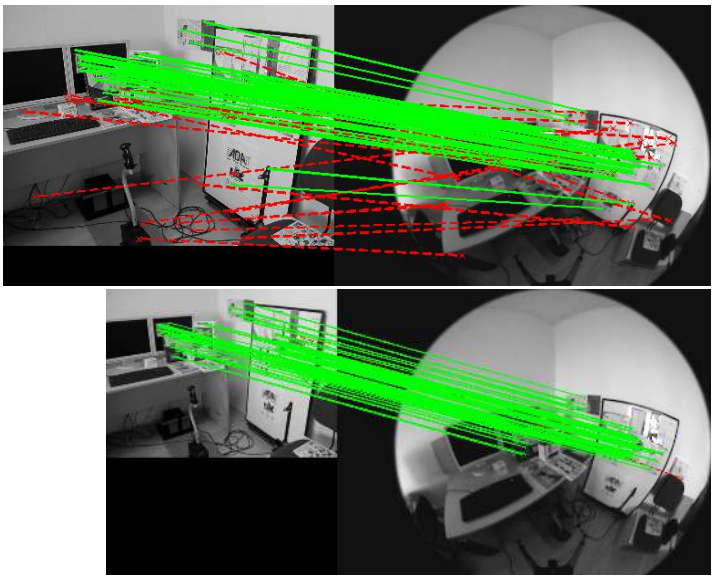


Figure 11: Matching results for the Pers2-Fish1 pair of the experiment in Fig. 10 for the direct SIFT matching (top) and matching after downsampling (bottom) approaches. Red dashed lines indicate false matches, whereas green lines indicate correct ones. Out of 75 matches, there are 21 false matches in the direct SIFT matching output, whereas there is only one false match with the proposed approach.

4. Epipolar Geometry and Motion Estimation

Epipolar geometry between hybrid camera views was explained by Sturm [9] for mixtures of paracatadioptric and perspective cameras. Barreto showed that the framework can also be extended to cameras with lens distortion due to the similarities between the paracatadioptric and division models [10]. To summarize this relationship, let us denote the corresponding image points in perspective and catadioptric images with $\mathbf{q}_p$ and $\mathbf{q}_c$ respectively. They are represented as 3-vectors in homogeneous 2D coordinates. To linearize the equations between catadioptric and perspective images, *lifted coordinates* are used for the points in omnidirectional images. Lifting for paracatadioptric cameras can be performed by $\hat{\mathbf{q}}_c = (x^2 + y^2, x, y, 1)^\mathsf{T}$. A 3x4 hybrid fundamental matrix expresses the epipolar constraint between these points:

$$\mathbf{q}_p^\mathsf{T} \mathsf{F}_{pc} \hat{\mathbf{q}}_c = 0 \tag{1}$$

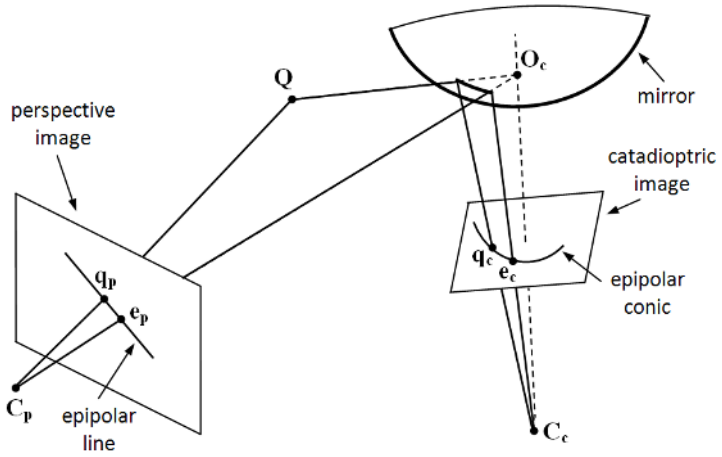


Figure 12: Hybrid epipolar geometry between a perspective and a catadioptric image. $\mathbf{q}_p$ and $\mathbf{q}_c$ are the projections of a 3D point $\mathbf{Q}$ on perspective and catadioptric images respectively. $\mathbf{e}_p$ and $\mathbf{e}_c$ are the epipoles in the perspective and catadioptric images respectively.

Using F_{pc}, geometric entity relations are:

$$\mathbf{l}_p = \mathsf{F}_{pc}\hat{\mathbf{q}}_c\ , \quad \mathbf{c}_c = \mathsf{F}_{pc}^{\mathsf{T}}\mathbf{q}_p\ , \quad \hat{\mathbf{q}}_c^{\mathsf{T}}\mathbf{c}_c = 0\ , \quad \mathbf{q}_p^{\mathsf{T}}\mathbf{l}_p = 0 \tag{2}$$

where $\mathbf{l}_p$ is the epipolar line in the perspective image and $\mathbf{c}_c$ is the corresponding epipolar curve (here, a circle) in the catadioptric image. Lifting coordinates enables us to represent a point on a curve with a simple dot product ($\hat{\mathbf{q}}_c^{\mathsf{T}}\mathbf{c}_c = 0$) as we do for line-point incidence. Actually $\mathbf{c}_c$ is a 4-vector containing the four different elements of a curve (epipolar conic in our case) represented in matrix form:

$$C = \begin{pmatrix} 2c_1 & 0 & c_2 \\ 0 & 2c_1 & c_3 \\ c_2 & c_3 & c_4 \end{pmatrix} \tag{3}$$

Hybrid epipolar geometry can be visualized in Fig. 12. An example of corresponding epipolar lines/conics are given in Fig. 13.

The relation between two paracatadioptric views can be represented by a 4x4 fundamental matrix. Lifted coordinates for hyperbolic mirrors are represented by 6-vectors since the corresponding conic does not have to be a circle. Hypercatadioptric images

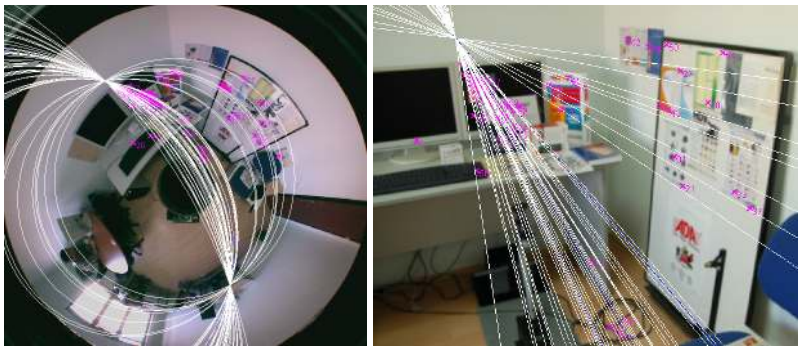


Figure 13: Example catadioptric-perspective pair and epipolar conics/lines of point correspondences.

fail to satisfy the same linear form of the epipolar constraint [10], however it has been shown that a linear relation exists with a 15x15 fundamental matrix for hypercatadioptric cameras [23].

4.1. Normalization

Normalization of image point coordinates comprises carrying the origin to the centroid of the points and scaling the coordinate values. It is crucial for fundamental matrix estimation as indicated in [37] for the 8-point algorithm of perspective cameras.

A way to perform normalization for lifted coordinates is normalizing point coordinates before lifting them. In this case, after F_{pc} is computed with lifted coordinates, we need to denormalize corresponding points, lines and conics for outlier elimination. However, we chose to define 4x4 T matrices for normalization of lifted coordinates ($\hat{\mathbf{q}}_{1norm} = \mathsf{T}_1\hat{\mathbf{q}}_1$ and $\hat{\mathbf{q}}_{2norm} = \mathsf{T}_2\hat{\mathbf{q}}_2$), so that normalized coordinates still suit to the lifted form, i.e. $(x^2 + y^2, x, y, 1)$.

Let n be the value of 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:

$$\hat{\mathbf{q}}_{norm} = \left(\frac{(x-c_x)^2}{n^2} + \frac{(y-c_y)^2}{n^2}, \quad \frac{x-c_x}{n}, \quad \frac{y-c_y}{n}, \quad 1 \right) \tag{4}$$

The transformation T defined in Eq. 5 yields $\hat{\mathbf{q}}_{norm}$ when multiplied with unnormalized lifted coordinates ($\hat{\mathbf{q}}$).

$$\hat{\mathbf{q}}_{norm} = \mathsf{T}\hat{\mathbf{q}} = \begin{pmatrix} \frac{1}{n^2} & \frac{-2c_x}{n^2} & \frac{-2c_y}{n^2} & \frac{c_x^2+c_y^2}{n^2} \\ 0 & \frac{1}{n} & 0 & \frac{-c_x}{n} \\ 0 & 0 & \frac{1}{n} & \frac{-c_y}{n} \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x^2+y^2 \\ x \\ y \\ 1 \end{pmatrix} \tag{5}$$

With these T matrices, denormalization of F can be performed linearly by $\mathsf{F} = \mathsf{T}_2^{\mathsf{T}}\mathsf{F}_{norm}\mathsf{T}_1$, which produces correct epipolar conics/lines (Eq. 2).

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 $\sqrt{2}$. It indicates the case that (x,y) coordinates are normalized to (1,1) which minimizes the difference between bare, multiplied and powered coordinate values [38]. This is the optimal condition for the linear estimation of F because these values are multiplied with the entries of F. We investigated the similar optimal case for hybrid pairs where we have a variety of terms such as $(x^2 + y^2)x$, xy, x and 1. Observing that no certain scale normalization value (n) equalizes all these terms as in the perspective camera case, we conducted experiments to find the empirically best value for hybrid pairs of real and simulated images [20]. We observed that $(n_{omni}, n_{pers}) = (\sqrt{2}, \sqrt{2})$ is the best performer for most of the cases.

4.2. Outlier Elimination

After the initial detection of the point matches, the RANSAC [13] algorithm based on the hybrid epipolar constraint is used to eliminate the false matches. Since F_{pc} has 12 elements, for RANSAC in hybrid images the minimum number of correspondences needed to estimate F_{pc} is 12-1(scale factor)=11. As in the perspective camera case, we define a distance threshold (d) to distinguish the outliers from inliers, where the points closer to their corresponding epipolar lines/curves 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. In

Image pairs	point matching		true/false after RANSAC
	total	true/false final	
Pers2 - Omni1	75	51/7	49.3/2.1
Pers2 σ1.5 d1.65 - Omni1	75	69/1	67.8/0.1
Pers2 σ1.5 - VCP1	75	73/1	70.7/0.0
Pers3 - Omni1	60	39/9	36.9/3.6
Pers3 σ1.5 d1.65 - Omni1	60	50/6	47.9/2.5
Pers3 σ1.5 - VCP1	60	57/1	54.5/0.0
Pers2 - Omni2	60	32/2	29.1/0.0
Pers2 σ2.5 d3.6 - Omni2	60	54/1	53.8/0.5
Pers2 σ2.5 - VCP2	60	56/3	54.6/0.0
Pers3 - Omni2	45	15/1	12.9/0.4
Pers3 σ2.5 d3.3 - Omni2	45	32/6	31.8/2.3
Pers3 σ2.5 - VCP2	45	35/3	33.4/1.0

Table 4: Matching results after RANSAC for the wide-baseline image pairs in Table 2. We repeated RANSAC 30 times for each pair and recorded the mean values yielding non-integer values. Distance threshold of RANSAC, d, was set to 15 pixels. Pers$N$ σA dB indicates that Pers$N$ image was blurred with $\sigma = A$ Gaussian filter and downsampled by a factor of B in each direction.

the VCP approach, we first calculate the coordinates of point correspondences in the omnidirectional image and then use them in RANSAC.

During F estimation, the rank 2 constraint is imposed by non-linear refinement of an orthonormal representation of F as proposed in [39] which was proved to provide better results than the direct imposition of rank 2.

Experiments. For the wide baseline image pairs of Fig. 6 and Table 2, the number of matches and successful match ratios before and after RANSAC elimination are given in Table 4. We infer from the table that the remaining false matches can be eliminated by RANSAC to a great extent. If there are still a few false matches it means these false matched points are very close to the corresponding epipolar line/conic by coincidence. The proposed point matching method (especially the VCP approach) provides a higher number of correct matches as input to RANSAC when compared to *Scale restricted SIFT* (rows with Pers$N$-Omni$M$). Since the final F is estimated by using all the inlier matches, more correct matches directly results in a more accurate estimate.

4.3. Motion Estimation

Motion estimation is the step of extracting the motion parameters of the cameras with respect to each other. The rotation and translation between the camera views are extracted from the essential matrix (E) with the technique given in [37]. We analyzed two methods for the estimation of E. The first option is directly estimating E with the calibrated 3D rays of the correspondences in the RANSAC output. The other option is estimating F_{pc} with RANSAC and then extracting E from F_{pc} using the relation [10]:

$$\underbrace{\mathbf{q}_p^{\mathsf{T}}\mathsf{K}_p^{-\mathsf{T}}}_{\bar{\mathbf{q}}_p^{\mathsf{T}}}\,\mathsf{E}\,\underbrace{\Theta^{\mathsf{T}}\hat{\mathsf{K}}_c^{\mathsf{T}}\hat{\mathbf{q}}_c}_{\bar{\mathbf{q}}_c} = 0 \tag{6}$$

where $\bar{\mathbf{q}}_p$ and $\bar{\mathbf{q}}_c$ are the normalized 3D rays for perspective and catadioptric cameras respectively. K_p is the calibration matrix of the perspective camera, $\hat{\mathsf{K}}_c$ is the lifted calibration matrix of the catadioptric camera. Finally

$$\Theta^{\mathsf{T}} = \begin{pmatrix} 0 & 2 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ -1 & 0 & 0 & 1 \end{pmatrix}$$

which carries the origin of coordinate system to the center of the sphere (cf. Fig. 2) linearly with lifted coordinates for paracatadioptric cameras.

Experiments. We conducted an experiment on simulated data to compare these two options. Fig. 14 shows the 3D location estimation errors and 2D (reprojection) errors for direct-E and E-from-F methods for varying number of point correspondences and varying amount of Gaussian location noise added to both images. We infer from the graphs that the direct-E method resulted in less 2D and 3D error. E-from-F is more vulnerable to noise. This is not surprising since lifted coordinates are used to compute F_{pc} which increases the impact of noise and F_{pc} has 12 elements whereas E has 9.

We should also keep in mind that the direct-E approach can be used for all types of cameras as long as calibration is performed, whereas the E-from-F approach is not practically possible for all omnidirectional cameras as the problems with hypercatadioptric systems are explained in [10].

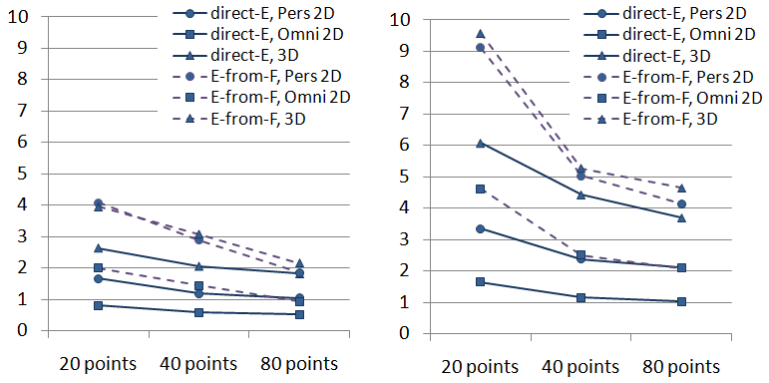


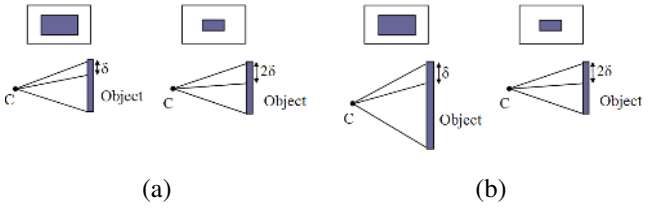
Figure 14: Comparison of the two E-estimation methods: direct-E and E-from-F. After E was estimated with either of these methods using point correspondences, 3D locations of the points were estimated with *linear-Eigen* triangulation and reprojected to the images. Plotted values are the 3D estimation errors (in cm.) and the median 2D reprojection errors (in pixels) for the images of hybrid pair. The number of the point correspondences used to compute E and F is given in x axis. Each case was repeated 50 times and mean values are plotted. Solid lines show the errors of direct-E method, dashed lines show errors of E-from-F method. (a) σ of the Gaussian noise is 1.0 pixel. (b) σ of the Gaussian noise is 2.0 pixels.

5. Triangulation

We propose an improvement for the *iterative linear-Eigen* triangulation method for effective use in hybrid SfM. According to the comprehensive study by Hartley and Sturm [17], *iterative linear-Eigen* is one of best triangulation methods for Euclidean reconstruction. It is superior to the midpoint method and non-iterative linear methods especially when 2D error is considered.

Let the two corresponding points be $\mathbf{q} = (x, y, 1)$, $\mathbf{q}' = (x', y', 1)$ which are obtained by projecting the 3D point $\mathbf{Q}$ on the images by $\mathbf{q} = \mathsf{P}\mathbf{Q}$, $\mathbf{q}' = \mathsf{P}'\mathbf{Q}$. Letting p_i denote the i^{th} row of P, *linear-Eigen* triangulation estimates $\mathbf{Q}$ by finding the least-squares solution of

$$\mathsf{A}\mathbf{Q} = 0 \quad \text{where} \quad \mathsf{A} = \begin{pmatrix} xp_3 - p_1 \\ yp_3 - p_2 \\ x'p'_3 - p'_1 \\ y'p'_3 - p'_2 \end{pmatrix} \tag{7}$$



(a) (b)

Figure 15: Depiction of doubling the focal length and decreasing the camera-scene distance for triangulation on normalized rays. On the top row, images are shown with the object observed. Bottom row shows the field of view of the camera (top and bottom 3D rays) corresponding to the object and the distance error on the object (δ or 2δ) corresponding to one pixel noise in the images. C represents the camera center (i.e. pinhole). (a) Camera on the left has twice the focal length of the camera on the right. Distance between the camera and the scene is equal for both cameras. One pixel noise in the zoomed image corresponds to lower angular error and distance error on the object. (b) Both cameras have equal focal length values but the distance between the camera on the left and the scene is half the distance between the camera on the right and the scene. Triangulation with the normalized 3D rays already gives support to left camera to minimize the angular error.

Two rows are added to A for each view. This method is extended by adjusting the weights of rows iteratively such that the reprojection error will be decreased resulting in *iterative linear-Eigen* triangulation. The weights for the first and the second views are $\frac{1}{p_3\mathbf{Q}}$ and $\frac{1}{p'_3\mathbf{Q}}$ respectively [17].

Please note that we employ this method with the 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 only be performed with the 3D rays outgoing from the effective viewpoints of the cameras.

The perspective cameras in hybrid systems tend to have higher resolution than the omnidirectional ones. To benefit from their resolution, we increased the weight of rows coming from the perspective images. With the mentioned weighting strategy, we observed improvement in the accuracy of the estimated 3D coordinates. We relate the amount of weighting to three factors: the scale ratio (r_s), the ratio of distances to the scene (r_d) and the position factor (p) as explained below. We propose to multiply the rows of the perspective image by s

$$s = r_s \cdot r_d \cdot p \tag{8}$$

In Fig. 15a, the case of doubling the focal length is depicted. If one pixel noise in the left image causes δ distance error, then the same amount of noise in the right image causes 2δ error. In this case, we need to increase the weight of the zoomed camera. We use the ratio of scales of the objects ($r_s = 2$).

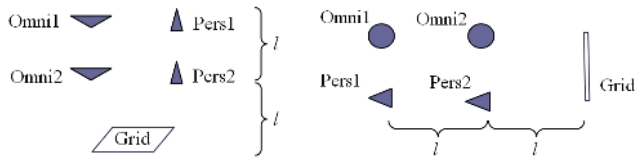
Distance to the scene affects the triangulation as well. The rays diverge as the camera comes closer to the scene. The object in the image gets larger but this is different from the zoom effect as depicted in Fig. 15b. Triangulation already gives support to the left camera because it minimizes the reprojection error on the normalized 3D rays. We should not increase the weight and we use r_d to compensate the zoom effect ($r_s = 2$, $r_d = 0.5$, $r_s \cdot r_d = 1$).

The third and the last factor is the position of the points in the omnidirectional image. When the objects in the scene have approximately the same height with the camera, x and y values in the $(x, y, 1)$ form of the normalized 3D rays have higher values compared to the values in the perspective images. These high values cause an unwanted support for the rows coming from the omnidirectional image. We observed that, for the points near the periphery of the omnidirectional image, increasing weight of perspective camera improves the results. The value of p is chosen empirically from our experiments. Detailed results of the experiments analyzing various cases are given in Section 5.1.

Please note that for two perspective cameras, iterative linear triangulation is performed on pixels (not on 3D rays) and since the reprojection error in the image is minimized, the zoomed image is supported without requiring an extra weighting. Distance effect works same as the zoom effect and position of the point in the image does not have a significant effect for perspective cameras.

5.1. Experiments with Simulated Images

We first analyze the improvement of the proposed weighted triangulation approach on a simulated environment. We generated a total of 1000 points which are regularly distributed on a planar grid and added Gaussian location noise to all simulated points. We define two main scenarios to distinguish between the cases when observed points are below or at the same horizontal level with the omnidirectional camera. We depicted



(a) (b)

Figure 16: Camera and grid positions in the scene of triangulation experiments. (a) The grid is below the catadioptric camera, side view. (b) The grid is at the same horizontal level with the cameras, top view. Distance between the cameras is 2 meters. Maximum distance between the cameras and the scene ($2l$) is 2.5 meters.

pair	r_s	r_d	$w = 1$	$w = s$	**improvement**
Omni1-Pers1	2	1	0.0250	0.0242	3.2%
Omni1-Pers1	4	1	0.0334	0.0302	9.6%
Omni1-Pers2	2	0.5	0.0200	0.0200	-
Omni2-Pers1	1	2	0.0170	0.0166	2.4%

Table 5: Results of triangulation experiments for Fig. 16a. For $w = 1$ and $w = s$ (proposed weighting) the values in the table are 3D coordinate estimation error in meters, median of 1000 points in the grid. Experiments were repeated 30 times and the values in the table are the mean of these 30 experiments. Gaussian location noise with σ=2.0 pixels was added to both images.

these two cases in Fig. 16. Two of the camera positions are selected each time also with varying focal length values to create the analyzed scale ratios.

For the case that the grid is below the omnidirectional camera, results are given in Table 5 where r_s and r_d are indicated as well. The applied weight value is represented with w. Errors for $w = 1$ and $w = s$ are compared in the table, where s is the proposed weighting computed by Eq. 8 and $p = 1$ for the current case. We express the improvement as percentage of decrease in error. Improvement becomes significant when r_s increases which is quite likely for hybrid pairs. For the Omni1-Pers2 pair, the effects of r_s and r_d cancel each other and $s = 1$ already. We included this case to indicate the importance of r_d because we tested that $w = 1$ is better than $w = 2$.

When the grid is close to the same horizontal level with the omnidirectional camera (Fig. 16b) we conducted the same experiments, results of which are given in Table 6. Applying the proposed weighting strategy decreases the 3D error by 2.3-9.6%. We

pair	r_s	r_d	p	$w=1$	$w=s$	**improvement**
Omni1-Pers1	2	1	2	0.0266	0.0254	4.5%
Omni1-Pers1	4	1	2	0.0317	0.0288	9.2%
Omni1-Pers2	2	0.5	2	0.0221	0.0216	2.3%
Omni2-Pers1	1	2	2	0.0171	0.0162	5.3%

Table 6: Results of triangulation experiments for Fig. 16b. Error is expressed by 3D coordinate estimation error in meters, median of 1000 points in the grid. Experiments were repeated 30 times and the values in the table are the mean of these 30 experiments. Gaussian location noise with σ=2.0 pixels was added to both images.

took $p = 2$ as it gave the best results in our experiments. When the observed scene points are below the horizontal level of the omnidirectional camera but not directly below (a case between Fig. 16a and Fig. 16b), it is appropriate to increase p from 1 to 2 gradually as the 3D points get higher. The value of p can be assigned according to the vertical angles of the 3D rays corresponding to the points as demonstrated in the experiment in Section 5.2.

5.2. Experiments with Real Images

Here, we combine the triangulation with the steps of point matching and epipolar geometry estimation to complete an SfM experiment with a real hybrid image pair. We analyze the improvement gained by the proposed weighted triangulation.

We use the RANSAC output for Pers2 σ1.5 - VCP1 pair (cf. Table 4), which has 70 correspondences. We estimate the essential matrix with the calibrated rays of these point correspondences. In Fig. 17, correspondences on the images and the reconstructed scene are given. 3D coordinates were computed with the proposed weighted *iterative linear-Eigen* triangulation, where Eq. 8 was used to compute weights. We take r_s=1.65 which was already extracted in the point matching step for the current image pair (Section 3.2, Table 2). We know that $r_d \approx 2$ since we set up the experiment environment, however one can also use the result of an initial triangulation (like Fig. 17 bottom row) to obtain an approximate ratio of distances to the scene. We employ varying p values for the points according to the vertical angles of their corresponding 3D rays. The vertical angles change between 50°-90° (0° indicates directly below the

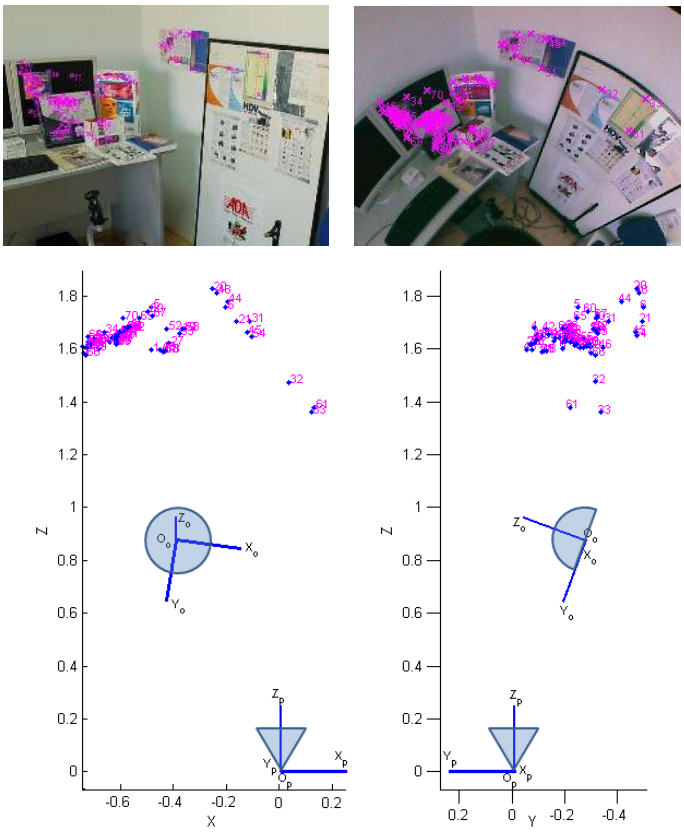


Figure 17: Reconstruction with a hybrid real image pair. Selected correspondences on images are viewed on top. Images were cropped to make the points distinguishable. At the bottom, 2D top-view (left) and 2D side-view (right) of the reconstructed scene can be observed. O_p and O_o shows the perspective and omnidirectional camera centers, $(\mathsf{X}_p, \mathsf{Y}_p, \mathsf{Z}_p)$ and $(\mathsf{X}_o, \mathsf{Y}_o, \mathsf{Z}_o)$ shows the perspective and omnidirectional camera axes, respectively. Plots were aligned w.r.t. the axes of the perspective camera. Actually, the optical axis of the omnidirectional camera (Z_o) is perpendicular to the floor and the perspective camera is looking slightly down.

omnidirectional camera) and we take p gradually increasing from 1.5 to 2 with the increasing angle.

To estimate the improvement gained by the proposed weighting scheme, we compare a number of real world distances with the ones in the estimated 3D structure. We measured 30 distances in the scene corresponding to the distances between estimated 3D points. They are not in the same scale, thus we equalized the scale of the distances

r_s	r_d	p	$w = 1$	$w = s$	improvement
1.65	2	1.5-2	0.88	0.83	5.7%

Table 7: Distance estimation errors after the triangulation for the hybrid real image pair. Error is expressed with the absolute difference between the measured real-world distances and the estimated distances after triangulation (in centimeters). Values are the median of 30 distance errors. For reference, these 30 measured distances vary between 11.2 cm. and 31.8 cm. with a median value of 16.5 cm.

using the ratio between the averages of 30 distances. We measure the accuracy with the absolute difference of the distances at the real scene and at the reconstructed scene. Table 7 shows the median of these 30 distance errors (in centimeters) for $w = 1$ and $w = s$. One can see the improvement brought by the proposed weighting scheme.

6. Adding Views and Bundle Adjustment

To integrate additional views for multi-view SfM, we employed the approach proposed by Beardsley *et al.*[18]. In this approach, when a sequence of views is available, initially SfM is applied for the first two views. Then, for each new view, feature matching is performed with the previous view and the features which correspond to the already reconstructed 3D points are detected. The projection matrix of the new view is computed using these final 2D-3D matches.

Sparse bundle adjustment (SBA), see e.g. the implementation of Lourakis and Argyros [19], has become popular in the community due to its capability of solving high dimensional 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. During bundle adjustment, 11 parameters are optimized for each view consisting of five intrinsic (ξ, focal length, aspect ratio, principal point coordinates), three rotation and three translation parameters.

7. Multi-view SfM Experiments

We first present a multi-view SfM experiment with real images of mixed cameras, thus we employ the entire pipeline shown in Fig. 1. All the views (two omnidirectional,

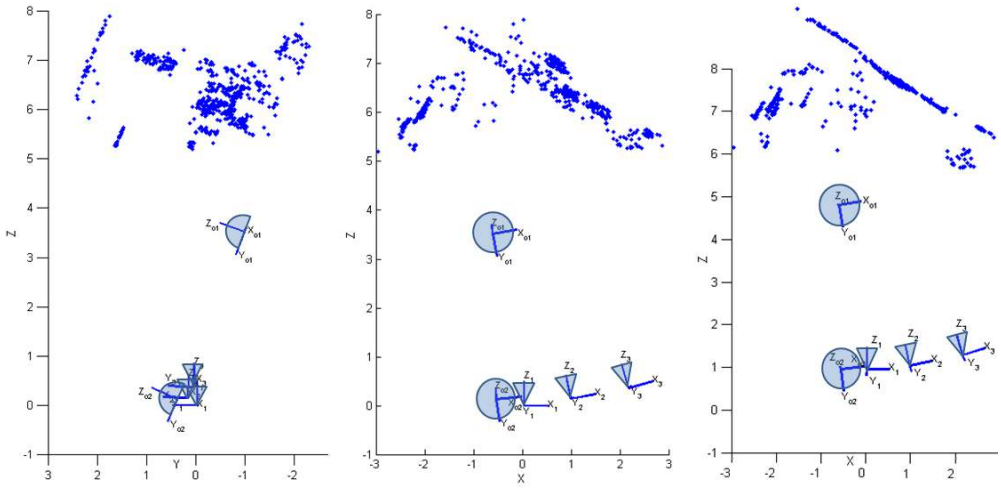


Figure 18: Estimated camera positions, orientations and scene points for the hybrid multi-view SfM experiment which includes the images in Fig. 6. Side-view is shown at left. At right and in the middle, two different top-views are shown. Plots at the left and in the middle are aligned w.r.t. the axes of the perspective camera no.1 (X_1, Y_1, Z_1). Actually, the optical axes of the omnidirectional cameras (Z_{oi}) are perpendicular to the floor and the perspective cameras are looking slightly down. The top-view at right is aligned with the omnidirectional cameras, therefore this view is perpendicular to the floor. It can be observed that the feature points on the wall and the clipboard, at the top of the point cloud, are correctly aligned.

three perspective) in Fig. 6 were used for this experiment. Estimated coordinates of the points (551 points were reconstructed) and estimated camera positions are shown in Fig. 18.

We performed SBA on this structure and camera parameters. The reprojection errors before and after SBA are given in Table 8 for all five views, where we observe that the errors were considerably decreased after SBA. The errors in the omnidirectional images are relatively higher than the ones in the perspective images. This is mainly due to the fact that the number of correspondences between the omnidirectional and perspective images is fewer than the number of correpondences between two perspective images. This decreases the accuracy of motion estimation for omnidirectional views as they are added to the structure. The table also shows the 3D errors (in cm.) before and after SBA. They were estimated in the same manner with the triangulation experiment given in Section 5.2 and Table 7. Same 30 real world distances were compared with

	Pers1	Pers2	Pers3	Omni1	Omni2	3D error
Before SBA	0.43	0.44	0.60	0.91	0.67	0.287
After SBA	0.23	0.19	0.24	0.39	0.47	0.268

Table 8: The mean values of reprojection error in images before and after SBA (in pixels) for the hybrid multi-view SfM experiment. 3D errors are the distance estimation errors (in cm.) and expressed with the absolute difference between the measured real-world distances and the estimates in the reconstruction. Same 30 real world distances and same method in Table 7 were used to compute distance errors.

the ones in the estimated 3D structure.

To demonstrate the effects of the bundle adjustment, we measured the 3D movement of the reconstructed points and the translation and rotation of the cameras between the initial reconstruction and the SBA result. The mean 3D movement for a point is 0.37cm. The average rotation in one axis is 0.3, 0.7, 1.9 and 3.6 degrees for cameras Pers2, Pers3, Omni1 and Omni2 respectively. Translation values are 2.1, 5.5, 14.4 and 7.8 cm. in the same order. Pers1 is kept fixed during the bundle adjustment and the rotation and translation values are estimated with respect to Pers1. We observe relatively higher rotation and translation adjustment for the omnidirectional cameras which were added to the structure with fewer common points. This is also true for Pers3 when compared to Pers2, since Pers3 was added to the structure only by the points common with both Pers1 and Pers2 by which the initial two view reconstruction was made.

7.1. Merging 3D Structures of Different Hybrid Image Pairs

Here, we discuss the theoretical and practical aspects of how an omnidirectional camera can combine the 3D structures viewed by two or more perspective cameras which do not have an overlapping view. This is often the case in a surveillance system scenario where an omnidirectional camera and limited number of fixed perspective cameras are present. The perspective cameras alone will not be adequate for SfM if their views do not overlap. Even when there is a small overlap, camera motion estimation becomes unreliable. Using omnidirectional cameras is advantageous since it reduces the number of perspective images to cover the gaps between the main images of interest.

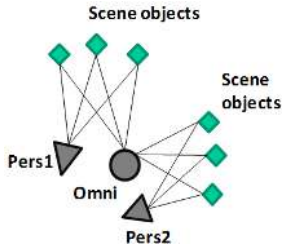


Figure 19: Depiction of merging 3D structures estimated with different hybrid image pairs.

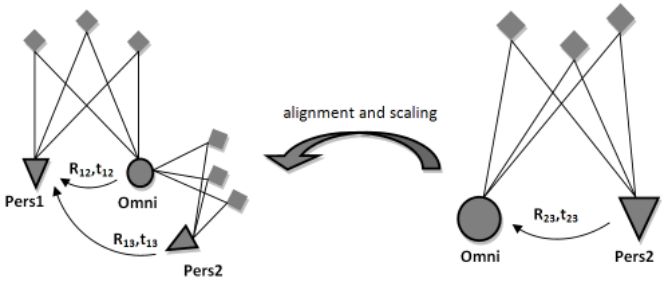


Figure 20: Depiction of aligning and scaling the 2^{nd} 3D structure w.r.t. the first one to obtain a combined structure.

A way of merging such perspective views is using more than one omnidirectional view. Some points are reconstructed only with the omnidirectional views to form a low resolution structure. Each perspective view can find common points with this structure and can be added to the structure by using the approach described in Section 6 and implemented at the beginning of this section.

Let us elaborate a more complicated case where only one omnidirectional view is used (Fig. 19). The multi-view approach does not work in this case since no common points can be reconstructed between perspective views. However, it is still possible to combine the 3D structures by pairing the perspective views with the single omnidirectional view. The two 3D structures obtained with different hybrid pairs are in different scales because the generated structures are up to a scale factor each. First, we have to align the two structures using the rotation and translation of the common view, then we have to adjust the scale (Fig. 20).

Let (R_{12},$\mathbf{t}_{12}$) be the rotation and translation between the first perspective image

and the omnidirectional image, and (R_{23},$\mathbf{t}_{23}$) be the ones between the omnidirectional image and the other perspective image. We need rotation and translation of the third camera (2^{nd} perspective image) with respect to the first camera, $\mathsf{K}_{ext,3} = [\mathsf{R}_{13} \,|\, \mathbf{t}_{13}]$, which can be computed as:

$$\mathsf{R}_{13} = \mathsf{R}_{23} \cdot \mathsf{R}_{12}$$

$$\mathbf{t}_{13} = \mathbf{t}_{23} + \mathsf{R}_{23} \cdot \mathbf{t}_{12}$$

To obtain the structure as a whole, we need to estimate the ratio of scales between two independently estimated structures and adjust the scale of the second structure by multiplying its translation vector ($\mathbf{t}_{13}$) by this ratio. In the following experiment, thanks to the small overlap between the perspective images, we use the 3D points which are available in both reconstructions to estimate the scale ratio. Since these points should be located at the same place in both reconstructions, the ratio of the distances between the points and the origin in the two reconstructions is the scale ratio. If there is no overlap, knowledge of real world distances in the scene or the distance between the cameras can be used to obtain the scale ratio.

In Fig. 21, on the top row, we observe the two perspective images, which have little overlap in their field of view. Out of 187 points, only four are common in all three images. These are the points that we align to obtain the scale ratio between the two structures. The estimated scale ratio is 0.334 in this experiment.

7.2. Integrating High Resolution and Low Resolution 3D Structures

In this section, we demonstrate that a high resolution (dense) and a low resolution (sparse) 3D structure can be obtained by a hybrid system for the purposes of foreground and background reconstruction respectively. The experiment we provide here serves as a proof-of-concept for the scenario in which omnidirectional cameras provide low resolution background reconstruction whereas the images of perspective cameras are used for modeling objects of interest in the foreground. In this way, the hybrid method relieves the need for a mobile camera or a network of cameras for background reconstruction.

The images of our experiment were taken in an everyday office environment. Two omnidirectional views were employed which is enough for a sparse background point

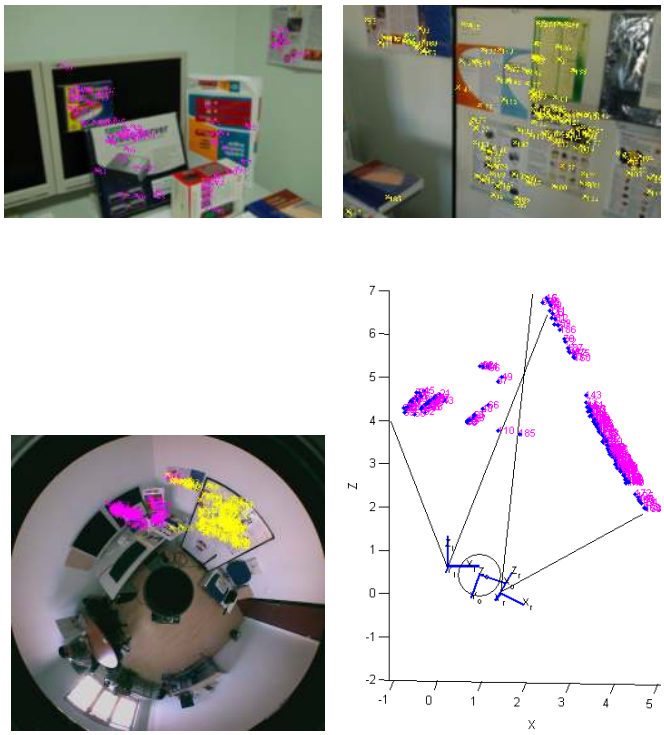


Figure 21: Correspondences between the perspective images (top row) and the omnidirectional image (bottom-left). Top-view of the estimated structure is given at bottom-right. Along with the reconstructed 3D points we also observe the positions, orientations and field of views of the cameras. The circle around the middle camera indicates the omnidirectional view.

reconstruction. Foreground object points were reconstructed by seven perspective views which resulted in a denser point cloud. In Fig. 22, sample perspective and omnidirectional images and reconstructed points can be seen. At bottom-left, dense structure (around 500 points) obtained by seven perspective images and estimated positions and orientations of perpective cameras are shown. When the sparse points reconstructed with omnidirectional views are added, we obtain the structure shown at bottom-right. There are a total of 614 points.

7.3. Increasing Perspective-SfM Accuracy with Omnidirectional Cameras

The wide field-of-view of omnidirectional cameras can increase the accuracy of camera motion estimation in a perspective camera SfM. In this section, we demonstrate

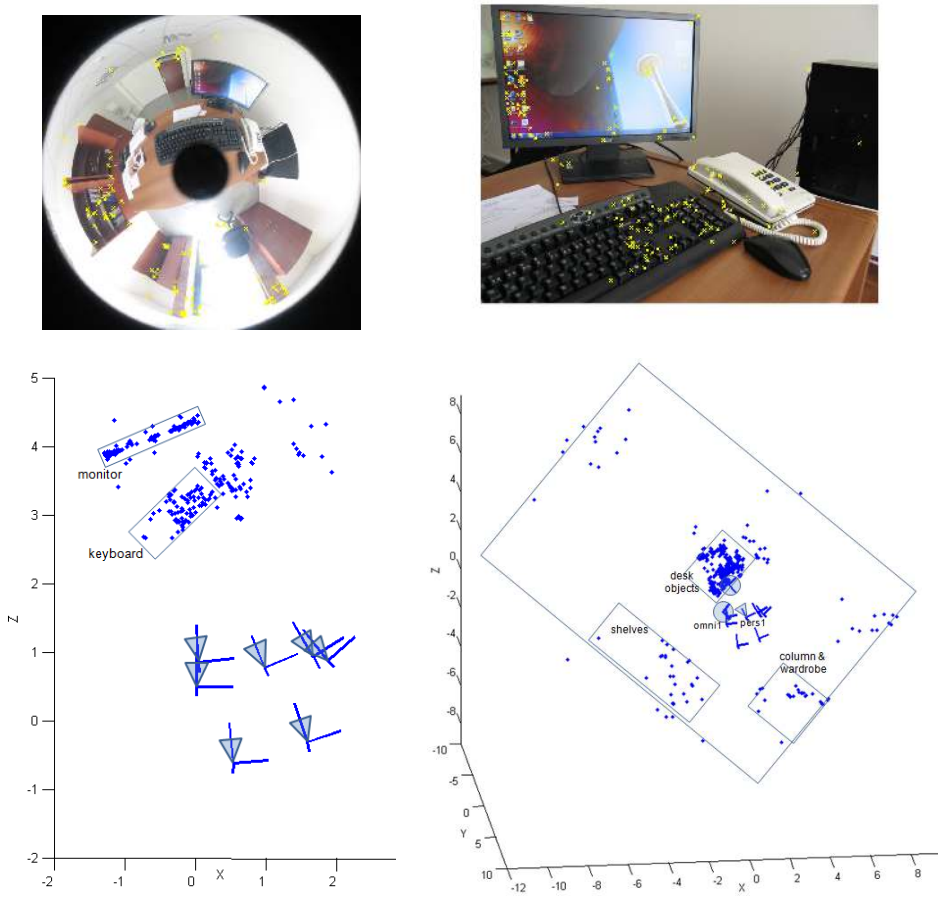


Figure 22: Sample perspective and omnidirectional images and reconstructed points for the experiment of low and high resolution 3D structure integration. At bottom-left, dense structure obtained by seven perspective images and estimated positions and orientations of perspective cameras are shown. At bottom-right, we see the whole structure containing the reconstructed background points and the dense point cloud belonging to the objects on the desk. There are a total of 614 points. Plots are aligned so that the optical axes of the omnidirectional cameras and also the viewer's looking direction are perpendicular to the floor. Objects in the room are indicated in the figure. Note that the location accuracy of the background objects is lower than that of the objects on the desk since only two omnidirectional images were used for background reconstruction.

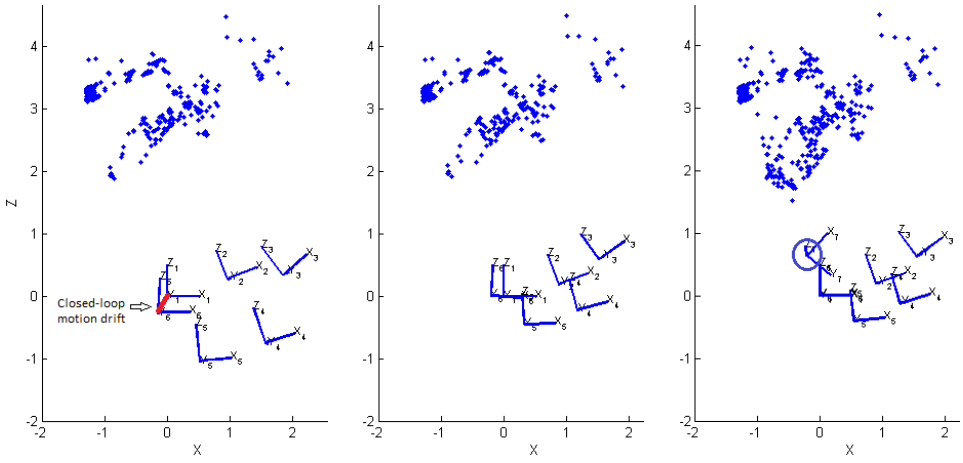
that adding an omnidirectional view considerably increases the accuracy of perspective-only structure and motion estimation.

In our experiment, we used a sequence of perspective images such that camera motion forms a closed loop. The scene is the same as the one used in Fig. 22. After the SfM pipeline was performed, the distance between the estimated positions of the first and the last cameras was measured. Since the actual positions of these two cameras are same, the measured distance indicates the drift in motion estimation. Fig. 23a shows the initial estimation of the structure and the positions of the cameras together with the drift. We compared this initial drift with the drift after applying bundle adjustment for both with and without adding the omnidirectional view. Fig. 23b shows the result for perspective-only case and Fig. 23c shows the result when the omnidirectional camera was added. We normalized the measured drift by dividing by the average of distances between all cameras in the loop. The initial drift was 20.55 cm, drift values after bundle adjustment were 18.05 cm and 0.95 cm for perspective-only SfM and hybrid SfM, respectively. This result clearly indicates that adding an omnidirectional camera considerably improves the accuracy of camera motion estimation. Approximately, one fifth of the scene points have a match in the omnidirectional view i.e. one fifth of the points were reconstructed with the omnidirectional view and at least one of the perspective views.

7.4. Outdoor SfM with an Image Sequence from a Video

To test the hybrid SfM in a more realistic scenario, we performed an outdoor experiment and used an image sequence from a captured video. More specifically, a frame per second was taken from a recorded video and a sequence of 18 frames was generated. A perspective SfM was performed which results in a point cloud (dense in certain locations in the scene). Then, this structure was enlarged with two omnidirectional views and point reconstructed with these two views which are from the regions that were not viewed by the perspective cameras. Fig. 24 shows sample perspective and omnidirectional images used in this experiment.

In Fig. 25, the result of the perspective SfM with the camera positions is shown at top-left. The estimated camera motion in particular is shown at a larger scale at



(a) (b) (c)

Figure 23: Results of the experiment to compare perspective-only SfM with hybrid SfM using the measured drift in a camera motion that forms a closed loop. The measured drift was normalized by dividing by the average of distances between all cameras in the loop. (a) The initial estimation of the structure and the positions of the cameras (viewed from the top). The drift depicted in the figure which has a value of 20.55 cm. (b) The result for perspective-only SfM after bundle adjustment. The drift decreased to 18.05 cm. (c) The structure and motion estimation after bundle adjustment when an omnidirectional view was added, which resulted in a drift value of 0.95 cm. The omnidirectional camera is indicated with a circle around it. Notice the increased number of points in the structure. There are nearly 500 reconstructed points and more than 100 of them were reconstructed with the omnidirectional view and at least one of the perspective views.

top-right. The motion is nearly linear and there are minor changes in camera orientations. When the omnidirectional views were added, we obtained the integrated structure shown in Fig. 25 bottom row. Dense parts were reconstructed using the perspective views and sparse points in different directions were reconstructed with the omnidirectional views.

8. Conclusions

We described an SfM pipeline and proposed new approaches or improved existing methods for the steps of this pipeline so that hybrid camera scenarios are covered.

It had been stated that directly applying SIFT is not sufficient to obtain good results for hybrid image pairs. In our study, we analyzed the reasons of false matches in SIFT

Figure 24: Sample images from the outdoor hybrid SfM experiment presented in Section 7.4.

and proposed a preprocessing algorithm that increases the matching performance considerably. After a few remaining false matches are eliminated by employing RANSAC on the hybrid epipolar constraint, we obtain a reliable set for motion estimation. Thus, we present automatic point matching between omnidirectional and perspective images.

We introduced the normalization matrices for lifted coordinates so that normalization and denormalization can be performed linearly. We evaluated the alternatives for motion estimation and decided on estimating the essential matrix with the calibrated 3D rays of point correspondences. We proposed a weighting strategy for iterative linear triangulation to improve the structure estimation accuracy and presented results with simulated and real data. Finally, we employed sparse bundle adjustment by adapting it to use the sphere camera model.

In conclusion, it is possible to perform hybrid multi-view SfM in an effective and automatic way. The usage of the sphere camera model throughout the pipeline was shown in this study. With the real image experiments, we showed that the proposed approach can be used effectively in the presence of hybrid camera systems. We handled two real world scenarios where employing a hybrid system is advantageous. One of them is a surveillance setup where the number of perspective cameras is limited and an omnidirectional camera can combine the 3D structures viewed by two or more perspective cameras which do not have an overlapping view. Another scenario is the 3D reconstruction in which the omnidirectional cameras provide low resolution background reconstruction whereas the images of perspective cameras are used for model-

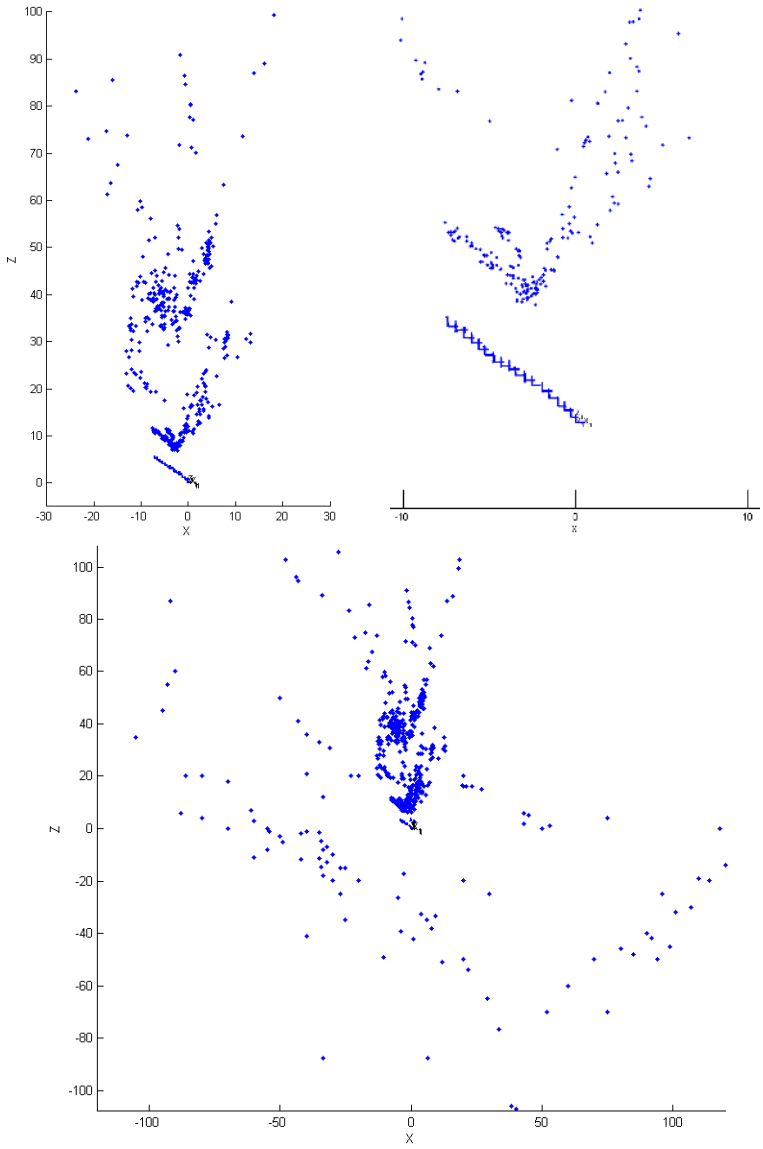


Figure 25: Integrated point cloud as a result of the outdoor hybrid SfM experiment. Samples of the images used in this experiment are shown in Fig. 24. Top-left: Estimated structure with perspective frames. Most of the points are from the border of the garden, plants in the garden and the facade of the house. Top-right: Estimated camera motion in particular at a larger scale. There are a total of 18 perspective views. Bottom: The integrated structure after the omnidirectional views were added. Dense parts were reconstructed using the perspective images and sparse points exist in different directions were reconstructed with the omnidirectional images. All the cameras are located at the center of the figure.

ing objects in the foreground. We also showed that adding omnidirectional cameras to perspective camera SfM increases the accuracy of camera motion estimation.

One of the future research directions may be developing an efficient hybrid dense depth estimation method and integrating it to the pipeline so that the framework can serve image-based 3D reconstruction applications.

Usage of fisheye camera images was limited to the point matching step in our study, since these cameras are not single-viewpoint systems and "ideally" can not be represented by the sphere model. On the other hand, it had been stated through empirical observations that the projection of some fisheye lenses can be approximated by the sphere camera model [40]. Such fisheye cameras can be used in the rest of the SfM pipeline as well.

References

[1] G. Scotti, L. Marcenaro, C. Coelho, F. Selvaggi, C. Regazzoni, Dual Camera Intelligent Sensor For High Definition 360 Degrees Surveillance, IEE Proc.-Vision Image Signal Processing 152(2) (2005) 250–257.

[2] K. Yamazawa, N. Yokoya, Detecting Moving Objects with an Omnidirectional Camera Based on Adaptive Background Subtraction, Lecture Notes in Computer Science 2749 (2003) 159–180.

[3] G. Adorni, S. Cagnoni, M. Mordonini, A. Sgorbissa, Omnidirectional Stereo Systems for Robot Navigation, in: Proc. of Workshop on Omnidirectional Vision (OMNIVIS), 2003.

[4] T. Goedeme, M. Nuttin, T. Tuytelaars, L. Gool, Omnidirectional Vision Based Topological Navigation, International Journal of Computer Vision (IJCV) 74(3) (2007) 219–236.

[5] S. Fleck, F. Busch, P. Biber, H. Andreasson, W. Strasser, Omnidirectional 3D Modeling on a Mobile Robot using Graph Cuts, in: Proc. of International Conference on Robotics and Automation (ICRA), 2005.

[6] M. Lhuillier, Toward Flexible 3D Modeling using a Catadioptric Camera, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[7] P. Chang, M. Hebert, Omni-directional Structure from Motion, in: Proc. of IEEE Workshop on Omnidirectional Vision, 2000.

[8] D. Chen, J. Yang, Image Registration with Uncalibrated Cameras in Hybrid Vision Systems, in: Workshop on the Applications of Computer Vision, 2005.

[9] P. Sturm, Mixing Catadioptric and Perspective Cameras, in: Proc. of Workshop on Omnidirectional Vision (OMNIVIS), 37–44, 2002.

[10] J. Barreto, K. Daniilidis, Epipolar Geometry of Central Projection Systems using Veronese Maps, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1258–1265, 2006.

[11] L. Puig, J. Guerrero, P. Sturm, Matching of Omnidirectional and Perspective Images using the Hydrid Fundamental Matrix, in: Proc. of Workshop on Omnidirectional Vision (OMNIVIS), 2008.

[12] D. Lowe, Distinctive Image Features From Scale Invariant Keypoints, International Journal of Computer Vision (IJCV) 60 (2004) 91–110.

[13] M. Fischler, R. Bolles, Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, Communications of the ACM 24(6) (1981) 381–395.

[14] S. Ramalingam, S. Lodha, P. Sturm, A Generic Structure-from-motion Algorithm for Cross-camera Scenarios, in: Proc. of IEEE Workshop on Omnidirectional Vision, 2004.

[15] C. Geyer, K. Daniilidis, A Unifying Theory for Central Panoramic Systems, in: Proc. of European Conference on Computer Vision (ECCV), 445–461, 2000.

[16] J. Matas, O. Chum, M. Urban, T. Pajdla, Robust Wide Baseline Stereo from Maximally Stable Extremal Regions, in: Proc. of British Machine Vision Conference (BMVC), 2002.

[17] R. Hartley, P. Sturm, Triangulation, Computer Vision and Image Understanding (CVIU) 68(2) (1997) 146–157.

[18] P. Beardsley, A. Zisserman, D. Murray, Sequential Updating of Projective and Affine Structure from Motion, International Journal of Computer Vision (IJCV) 23(3) (1997) 235–259.

[19] M. Lourakis, 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.

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

[21] Y. Bastanlar, A. Temizel, Y. Yardimci, P. Sturm, Effective Structure-from-Motion for Hybrid Camera Systems, in: Proc. of International Conference on Pattern Recognition (ICPR), 2010.

[22] J. Barreto, K. Daniilidis, Unifying Liftings for Catadioptric and Dioptric Cameras, in: Proc. of Workshop on Omnidirectional Vision, 2004.

[23] P. Sturm, J. Barreto, General Imaging Geometry for Central Catadioptric Cameras, in: Proc. of European Conference on Computer Vision (ECCV), 2008.

[24] C. Geyer, K. Daniilidis, Paracatadioptric Camera Calibration, IEEE Transactions on Pattern Analysis and Machine Intelligence 24(5) (2002) 687–695.

[25] S. Kang, Catadioptric Self-calibration, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 1201–1207, 2000.

[26] R. Orghidan, J. Salvi, M. Mouaddib, Calibration of a Structured Light-based Stereo Catadioptric Sensor, in: Proc. of Workshop on Omnidirectional Vision (OMNIVIS), 2003.

[27] C. Cauchois, E. Brassart, C. Drocourt, Calibration of the Omnidirectional Vision Sensor: SYCLOP, in: Proc. of International Conference on Robotics and Automation (ICRA), 1999.

[28] J. Kannala, S. Brandt, A Generic Camera Calibration Method for Fish-eye Lenses, in: Proc. of International Conference on Pattern Recognition (ICPR), 2004.

[29] D. Scaramuzza, A. Martinelli, R. Siegwart, A Toolbox for Easily Calibrating Omnidirectional Cameras, in: Proc. of Int. Conference on Intelligent Robots and Systems (IROS), 2006.

[30] J. Tardif, P. Sturm, S. Roy, Self-calibration of a General Radially Symmetric Distortion Model, in: Proc. of European Conference on Computer Vision (ECCV), 2006.

[31] C. Mei, P. Rives, Single Viewpoint Omnidirectional Camera Calibration from Planar Grids, in: Proc. of International Conference on Pattern Recognition (ICPR), 3945–3950, 2007.

[32] Y. Bastanlar, L. Puig, P. Sturm, J. Guerrero, J. Barreto, DLT-like Calibration of Central Catadioptric Cameras, in: Proc. of Workshop on Omnidirectional Vision, 2008.

[33] Z. Yi, C. Zhiguo, X. Yang, Multi-spectral Remote Image Registration based on SIFT, Electronics Letters 44(2) (2008) 107–108.

[34] F. Alhwarin, C. Wang, D. Ristic-Durrant, A. Gräser, Improved SIFT-Features Matching for Object Recongnition, in: Visions of Computer Science - BCS International Academic Conference, 2008.

[35] Y. Bastanlar, A. Temizel, Y. Yardimci, Improved SIFT matching for image pairs with scale difference, Electronics Letters 46(5) (2010) 346–348.

[36] V. Peri, S. Nayar, Generation of Perspective and Panoramic Video from Omnidirectional Video, in: Proc. of DARPA Image Understanding Workshop, 1997.

[37] R. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge Univ. Press, 2nd edn., 2004.

[38] R. Hartley, In Defense of the Eight-Point Algorithm, IEEE Transactions on Pattern Analysis and Machine Intelligence 19(6) (1997) 580–593.

[39] A. Bartoli, P. Sturm, Non-linear Estimation of the Fundamental Matrix with Minimal Paremeters, IEEE Transactions on Pattern Analysis and Machine Intelligence 26(3) (2004) 426–432.

[40] X. Ying, Z. Hu, Can We Consider Central Catadioptric Cameras and Fisheye Cameras Within a Unified Imaging Model?, in: Proc. of European Conference on Computer Vision (ECCV), 442–455, 2004.