Structure from motion using a hybrid stereo-vision system

François Rameau, Désiré Sidibé, Cédric Demonceaux, David Fofi

The 12th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI 2015)
October 28 ∼ 30, 2015 / KINTEX, Goyang city, Korea

Structure from motion using a hybrid stereo-vision system

François Rameau, Désiré Sidibé, Cédric Demonceaux, and David Fofi

Université de Bourgogne, Le2i UMR 5158 CNRS, 12 rue de la fonderie, 71200 Le Creusot, France

Abstract - This paper is dedicated to robotic navigation using an original hybrid-vision setup combining the advantages offered by two different types of camera. This couple of cameras is composed of one perspective camera associated with one fisheye camera. This kind of configuration, is also known under the name of foveated vision system since it is inspired by the human vision system and allows both a wide field of view and a detail front view of the scene.
Here, we propose a generic and robust approach for SFM, which is compatible with a very broad spectrum of multi-camera vision systems, suitable for perspective and omnidirectional cameras, with or without overlapping field of view.

Keywords - Hybrid vision, SFM

1. Introduction

Binocular vision system is a well-known configuration in computer vision which has been studied over the past decades. This configuration of two similar cameras is widely used for 3D reconstruction, mobile robot navigation, etc. Such type of system is particularly interesting because it allows the simultaneous capture of two akin images from which stereo matching can be achieved accurately using geometrical constraints and photometric descriptors.
In this paper we propose to modify the conventional stereo-vision system by replacing one of the cameras by an omnidirectional sensor, more specifically a fisheye camera. This new combination of cameras is very versatile as it combines the advantages from both cameras, offering desirable features for robot localisation and mapping. Indeed, the fisheye camera provides a large vision of the scene. Furthermore, it has been proved in [1] that spherical sensors are an appropriate solution to overcome the ambiguities when small amplitude motions are performed. On the other hand, the perspective camera can be employed to extract details from the scene captured by the sensor. The other advantage offered by this second camera is the possibility to estimate the displacements of a robot at the real scale.
In this paper, we propose a novel approach to estimate the motion of a mobile robot using this configuration of cameras in an efficient way. This approach is mainly inspired by non-overlapping SFM techniques.

This article is organized in the following manner. In the next section we give a definition of the term hybrid vision system and review the previous works in 3D reconstruction using hybrid vision system. In the section 3we describe our SFM framework adapted for heterogeneous vision system which does not need inter-camera correspondences. While the third part of this paper (section 4.) concerns the results obtained with our method. Finally, a short conclusion ends this article.

This work was supported by DGA (Direction Generale de l'Armement) and the regional council of Burgundy.

2. Previous works

In this section we review the already existing methods developed for the calibration and the navigation using hybrid-vision system. We are also giving a clear definition about the term "hybrid-vision system" and their uses.

2.1 Hybrid vision system

The term of hybrid vision system means that the cameras used within the vision system are of different natures or modalities [2]. This type of camera association allows the acquisition of complementary informations, for instance, an extension of the field of view, depth information or the study of a wider range of wavelength.
For example, in [3] the authors proposed to merge information from a conventional binocular system and from two infrared cameras in order to improve pedestrian detection process.
RGB-D sensors, such as the Kinect, using a reconstruction approach based on the projection of an infra-red pattern can also be viewed as hybrid vision sensor. In fact, a RGB camera is used to texture the reconstruction with visible color, while the infra-red camera can analyse the pattern to reconstruct the 3D structure of the scene. These two sensors are very complementary. For instance, in [4] the registration of 3D point cloud is simplified by the utilization of information from the RGB images. Many others original combination exist, for example in [5], where the sensor consists in the association of a high resolution camera with two low resolution cameras for real time events detection.
In this article, we are interested in the case where cameras have different geometrical properties. More explicitly, the association of one perspective and one omnidirectional camera. This specific type of system has already been studied especially for video-surveillance purposes for its ability to obtain conjointly a global view of the scene and an accurate image of the target from one or multiple perspective cameras. The calibration of such type of system is discussed in [6] where one omnidirectional camera is used in collaboration with a network of perspective cameras.
This configuration of camera can potentially be valuable for robotic navigation. In [7], the authors propose to use a catadioptric sensor and a perspective camera for obsta-

cle detection. For multi-robot collaboration, Roberti *et al.* [8] propose an original approach to compute the structure and the motion from multiple robots equipped with different types of camera.
Eynard *et al.* [2] also take advantage of this setup in order to estimate both the attitude and the altitude of a UAV using a dense plane sweeping based registration.

2.2 3D reconstruction and localisation using a hybrid vision system

Classical structure from motion methods consist in the joint estimation of the scene structure and the motion from a single moving camera [9]. When this type of approach is used to determine the displacements of a robot, the images are sorted temporally. Consequently, they are processed one after the other at every new frame acquisition, we call this strategy a sequential SFM. The 3D reconstruction as well as the motion estimation obtained from these methods are up to a scale factor, which represents a limitation in robotic navigation. To overcome this specific problem, a common solution is to utilize multiple cameras, the most basic example being the conventional stereo-vision system (usually two similar perspective cameras). Nevertheless, this sort of configuration needs a full calibration of the system. Furthermore, an accurate synchronisation of the cameras is mandatory in order to ensure a simultaneous images acquisition. The stereo image matching can be used to estimate the reconstruction of the environment with real scale. Most of the approaches of SFM using stereo-vision systems can be split into two main steps, the first one being the 3D reconstruction of the environment at a time t using stereo-correspondences only. The second step is the temporal tracking of the feature points in the next images. A pose estimation can be performed by minimizing the re-projection error of the 3D points on the new images acquired at time $t + 1$ [10].

Other techniques are also possible. In [11] the authors proposed a motion estimation using two interlaced trifocal tensors in order to estimate the six degrees of freedom of the stereo-rig motion. A more sophisticated way presented in [12] uses quadrifocal tensor in order to recover the motion parameters by dense image registration. These approaches are very efficient as illustrated by the results obtained with the KITTI dataset [13], however the majority of them concerns perspective cameras only. Only few works focused on visual odometry based on a hybrid stereo vision system. In [2], the displacement of a UAV is evaluated from a hybrid vision system but it is limited to aerial application since the main assumption is the planarity of the observed surface (the ground). To the best of our knowledge, no methods have been designed for the particular case of calibrated hybrid stereo-vision SFM. The main difficulty arising with these specific types of system equipped with one omnidirectionnal camera and one perspective camera, is the stereo matching between two images of different nature. Multiple articles tackled the problem of hybrid image correspondence, most of the current approaches are based on the adaptation of the usual descriptor to the geometry of the cameras (for instance Harris [14] or SIFT [15]). These descriptors are often used in conjunction with appropriated geometric constraints in order to remove outliers. The mentioned approaches can be used to achieve 3D reconstruction of the scene and to localize the cameras, however, they do not consider a calibrated stereo-vision system. In fact, this prior calibration is carrying valuable informations which can simplify the images matching, for instance, by rectification. Nevertheless, this epipolar rectification with cameras having notable different resolutions does not allow an accurate matching. Clearly, the inter-camera matching is a particularly complicated step and needs the use of sophisticated and computationally expensive approaches. Moreover, the accuracy offered by such type of process is highly depending upon the dissimilarity between the cameras resolution. Indeed, this difference is emphasized as the focal length of the perspective camera increases, making the previously described methods inefficient.
Nevertheless, the point matching between cameras of same nature (omnidirectional or perspective) is a relatively basic process since the usual descriptors remains very efficient. The proposed method takes advantage of this by getting rid of stereoscopic matching using a non overlapping field of view SFM method.

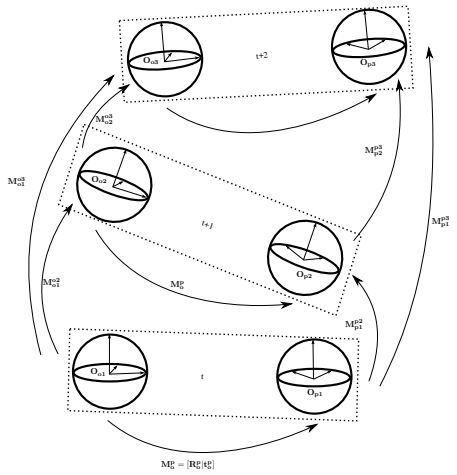


Fig. 1 Model of our system for two successive displacements

3. Methodology

In this paper we propose an adaptation of the non-overlapping SFM method develloped by Clipp *et al.* [16] to our specific case, that is to say a hybrid vision system in motion as seen in figure 1. The original method is very sensitive to degenerated motion, so we propose to use a new formulation of the problem based on tri-focal tensor in order to robustify the approach.

3.1 The hybrid stereo-vision system in motion

In this section, we analyse the multi-views relationships of our system. In a first place we consider two

successive displacements of the rig at times t, $t+1$ and $t+2$. Hence, our system is fixed and calibrated, it means that the inter-cameras transformation between the fisheye camera (o) and the perspective camera (p) written $\boldsymbol{M}_o^p = \begin{pmatrix} \boldsymbol{R}_o^p & \boldsymbol{t}_o^p \\ 0 & 1 \end{pmatrix}$ is known.
The displacements of the cameras are then linked by the following relations:

$$\boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o2} (\boldsymbol{M}_o^p)^{-1} = \boldsymbol{M}_{p1}^{p2} \quad (1)$$

$$\boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o3} (\boldsymbol{M}_o^p)^{-1} = \boldsymbol{M}_{p1}^{p3} \quad (2)$$

This rigid transformation between our cameras reduced the number of degrees of freedom to 6 in case of a single motion, as it is the case in [16]. This number of DOF rises to 11 for two motions of our vision system, this is the scenario which is examined in this paper.
Despite an overlapping field of view, our approach does not consider any stereo correspondences between the two cameras. In other terms, only temporal correspondences are employed.
The only required conditions for our approach are the detection of 6 triplet of corresponding points on one camera and 1 triplet on the other one.
Indeed, our approach is essentially based on the fact that it is possible to estimate the displacement of one camera from the computation of a trifocal tensor using 6 temporal correspondences [17]. Another minimal solution for calibrated cameras has been proposed by Nister and Schaffalitzky in [18], but this approach is less robust and computationally more complex. Projection matrices can therefore be extracted from the mentioned tri-focal tensor using the approaches described in [17]. Like the essential matrix, this estimation is done up to a scale factor, leaving only a single degree of freedom to estimate. This can be solved with a single matching triplet on the other camera constituting the vision system. Finally, the minimal solution needs 7 triplets of points in order to solve the motion and its scale together. Note that this can easily be extended for multi-camera setup with more than two cameras. Furthermore, it is compatible with any SVP camera thanks to the use of the unified projection model.

3.2 Estimation of the scale factor

The estimation of the two first displacements of the fisheye camera from the trifocal tensor gives $\boldsymbol{M}_{o1}^{o2}(\lambda)$ and $\boldsymbol{M}_{o1}^{o3}(\lambda)$ where λ is a unknown scale factor.
In this section, we describe an approach to retrieve this scale factor λ using only one triplet of corresponding points on the perspective camera.
The trifocal tensor $\boldsymbol{T}_p^{123}$ linking the three perspective views $p1$, $p2$ and $p3$ can be expressed as follow:

$$\boldsymbol{T}_p^{123} = [\boldsymbol{T}_{p1}, \boldsymbol{T}_{p2}, \boldsymbol{T}_{p3}] \quad (3)$$

$$\boldsymbol{T}_{p1} = \boldsymbol{M}_{p1}^{p2}[({}^2\boldsymbol{M}_{p1}^{p1})^T . {}^3\boldsymbol{M}_{p1}^{p1} - {}^3\boldsymbol{M}_{p1}^{p1} . {}^2\boldsymbol{M}_{p1}^{p1}]_{[\times]} (\boldsymbol{M}_{p1}^{p3})^T \quad (4)$$

$$\boldsymbol{T}_{p2} = \boldsymbol{M}_{p1}^{p2}[({}^3\boldsymbol{M}_{p1}^{p1})^T . {}^1\boldsymbol{M}_{p1}^{p1} - {}^1\boldsymbol{M}_{p1}^{p1} . {}^3\boldsymbol{M}_{p1}^{p1}]_{[\times]} (\boldsymbol{M}_{p1}^{p3})^T \quad (5)$$

$$\boldsymbol{T}_{p3} = \boldsymbol{M}_{p1}^{p2}[({}^1\boldsymbol{M}_{p1}^{p1})^T . {}^2\boldsymbol{M}_{p1}^{p1} - {}^2\boldsymbol{M}_{p1}^{p1} . {}^1\boldsymbol{M}_{p1}^{p1}]_{[\times]} (\boldsymbol{M}_{p1}^{p3})^T \quad (6)$$

where ${}^i\boldsymbol{M}$ is the i^{th} line of the matrix $\boldsymbol{M}$. The relations (1) lead to the following equations, which will be used to rewrite the perspective camera trifocal tensor only using the projection matrices of the omnidirectional camera:

$$\boldsymbol{M}_{p1}^{p1} = \boldsymbol{M}_o^p, \quad (7)$$

$$\boldsymbol{M}_{p1}^{p2} = \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o2}(\lambda), \quad (8)$$

$$\boldsymbol{M}_{p1}^{p3} = \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o3}(\lambda). \quad (9)$$

Then the tensor can be re-written:

$$\boldsymbol{T}_p^{123} = [\boldsymbol{T}_{p1}, \boldsymbol{T}_{p2}, \boldsymbol{T}_{p3}] \quad (10)$$

$$\boldsymbol{T}_{p1} = \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o2}(\lambda)[({}^2\boldsymbol{M}_o^p)^T . {}^3\boldsymbol{M}_o^p - {}^3\boldsymbol{M}_o^p . {}^2\boldsymbol{M}_o^p]_{[\times]} \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o3}(\lambda), \quad (11)$$

$$\boldsymbol{T}_{p2} = \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o2}(\lambda)[({}^3\boldsymbol{M}_o^p)^T . {}^1\boldsymbol{M}_o^p - {}^1\boldsymbol{M}_o^p . {}^3\boldsymbol{M}_o^p]_{[\times]} \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o3}(\lambda), \quad (12)$$

$$\boldsymbol{T}_{p3} = \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o2}(\lambda)[({}^1\boldsymbol{M}_o^p)^T . {}^2\boldsymbol{M}_o^p - {}^2\boldsymbol{M}_o^p . {}^1\boldsymbol{M}_o^p]_{[\times]} \boldsymbol{M}_o^p \boldsymbol{M}_{o1}^{o3}(\lambda). \quad (13)$$

Now, all the entries of the tensor are known, except the scale factor λ.
The point-point-point ($\boldsymbol{P}_{p1} - \boldsymbol{P}_{p2} - \boldsymbol{P}_{p3}$) transfer function validating this tensor is usually expressed as follow:

$$\boldsymbol{P}_{p2[\times]} (\sum_{i=1}^{3} \boldsymbol{P}_{p1}^i \boldsymbol{T}_{pi}) \boldsymbol{P}_{p3[\times]} = \boldsymbol{0}_{3\times 3}. \quad (14)$$

The unknown scale factor is embedded in the trifocal tensor $\boldsymbol{T}_{pi}$, every single triplet of point provide 9 dependant linear equations. Only one of these equation can be use to solve λ. Due to the complexity of the resulting equations, we obtained it using a computer algebra system (Matlab symbolic solver).

3.3 The algorithm

In a first step, the detected points over the three successive fisheye and perspective views are projected on the unitary sphere.
Thereafter, six triplets of points from the fisheye camera are used to compute a trifocal tensor. The method used in this work is the one described in [17]. The robust estimation of this tensor is provided by a RANSAC algorithm in order to remove outliers from our data. The reprojection error on the spheres being the criterion used to reject outliers in RANSAC.
The poses of the fisheye camera can be thereafter extracted from the tensor using methods from [17].At this point we have all poses up to a scale factor.

In order to determined the scale factor using only one triplet of point from the other camera, we used the equation obtained from the point-point-point transfer (14). Once again, a 1 point RANSAC is utilized in order to have a robust estimation of the scale factor.
Once all the poses of our cameras are retrieved, the linear solution obtained can be refined through an *ad-hoc* bundle adjustment process to ensure sufficient accuracy.

4. Results

4.1 Synthetic data experiments

To assess the quality and relevance of our approach dedicated to poses estimation for not overlapping field of view cameras, we propose in this section a series of synthetic tests. Our method using the trifocal tensor formalism is compared to [16], where the author develops a

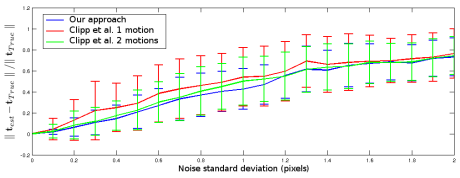


Fig. 2 Synthetic results with two camera for 100 iterations per noise level

method based on the estimation of the essential matrix.
The synthetic environment consists of a 3D point cloud generated randomly within a cube dimensions $500 \times 500 \times 500$, the cameras of the stereo-vision system are spaced by a distance of 20 units. The 3D point are projected on the image plan of both cameras, the fisheye camera is modelled using the spherical model and as a $180°$ horizontal, as well as, vertical field of view. The second camera is based on the pinhole model with a restricted field of view of $45°$. With such configuration, our simulation is very close to real case scenario.
The hybrid vision system is initially located in the center of the cube. For every new iteration of our test procedures, two new motions of the stereo rig are randomly generated. At each motion the rotation matrix is randomly modified over its 3 axes or rotation within a range of $\pm 6°$ while the translation is also changed between ± 10 units. In these experiments only non-degenerated cases are taking into account.
In order to test the robustness of our approach, a white Gaussian noise is adding on the coordinates of the images points. One hundred iterations are performed for each noise level.
The figure 2 shows the results obtained with our method compared with the one computed using the Clipp *et al.* approach for one and two pairs of views. The metric used for the comparison is the same proposed by the authors of the aforementioned method, it is $\| \boldsymbol{t}_{est} - \boldsymbol{t}_{True} \| / \| \boldsymbol{t}_{True} \|$ with $\boldsymbol{t}_{True}$ the ground truth and $\boldsymbol{t}_{est}$ the estimated translation of the perspective camera. This measurement offers the advantage to evaluate both the accuracy of the scale estimation and the direction of the translation. It is noticeable that our approach is slightly more robust than the one previously proposed method.

4.2 Experimentations with our system

In this section we present the experimental results obtained with our hybrid stereo-vision system composed of one perspective camera and one fisheye camera. For this assessment we are using two IDS μEye cameras with a spatial resolution of 1280×1024p. One of these cameras is equipped with a fisheye lens leading to a full hemispherical field of view (180°). On the second one, a more conventional lens is used which does not induce significant distortions in the image. Finally, these two cameras are rigidly mounted together on a rig (see figure 3). The tests were performed indoor in order to have a known measurement of the room that will be used as ground truth to validate the scale estimation. In the proposed sequence the cameras rig performed large and non-degenerated motions. The points detector and descriptor used is SURF for the temporal matching of fisheye and perspective images.

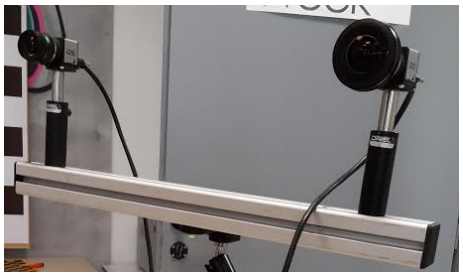


Fig. 3 Hybrid stereo-vision rig used for our evaluation

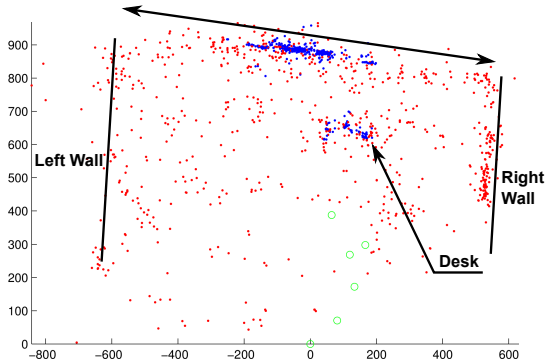
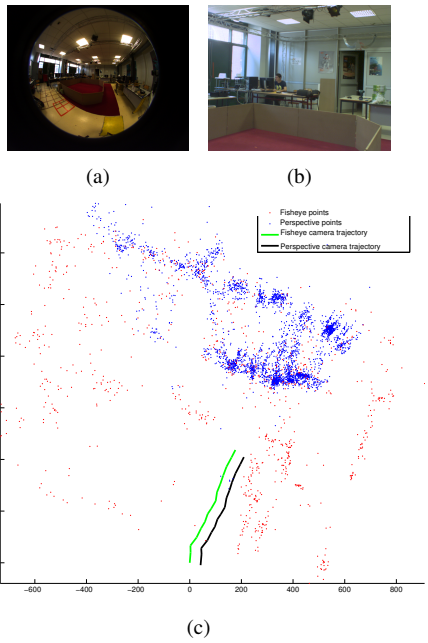


Fig. 4 3D reconstruction (upper view), the red points are computed from fisheye images while blue points are from the perspective camera. The green circles depicted the positions of the fisheye camera.

The figure 4 shows the 3D reconstruction of the environment computed with 6 views. On this figure, the red points are the 3D structure calculated from the fisheye views, in this case the reconstruction is sparse but covers a large area. The reconstruction from the perspective views is displayed in blue, we note that this reconstruction corresponds to a small portion of the scene viewed by the fisheye camera.
A prior measurement of the room gives a width of 10 meters, the width computed with our hybrid stereo vision system has the same order of magnitude, we can deduce that our poses estimation with a real scale factor is valid.

The figure 5 is another experimentation with 20 pairs of images. In the proposed reconstruction, the sparse point cloud computed from the fisheye camera has about 1000 3D points, however it is covering almost the whole room. On the other hand, there is a high density of points reconstructed using the perspective camera (more than 3000 points) on a smaller area.
This example highlighted the great advantage offered by our hybrid sensor, since a standard binocular system - despite their great precision- does not reconstruct such a wide area with a limited number of movements. Furthermore, a vision system strictly omnidirectional (for instance two catadioptric or fisheye cameras) is capable to reconstruct the scene entirely but without the accuracy and density offered by our perspective camera.



(a) (b) (c)

Fig. 5 Result obtained with 20 images per camera, (a) Fisheye image from the sequence, (b) Perspective image from the sequence, (g) 3D reconstruction computed

4.3 Tests with the KITTI database

Without an accurate and trustful ground truth it is impossible to provide a pertinent evaluation of our algorithm, this is the reason why we propose multiples tests done with the KITTI dataset [1].
These freely available data contains information from a large number of sensors, such as, monochromatic and color stereo vision systems, a LIDAR, a IMU and a GPS. Theses sensors are mounted on a vehicle travelling in the streets of Karlsruhe (Germany). The KITTI dataset for visual odometry contains 22 sequences of different type and length from few hundred meters up to multiple kilometres. An accurate ground truth (localisation error inferior to 10cm) is also provided.
For our experimentations we use the greyscale images from two 1.4 Megapixels resolution Grey Flea 2 cameras. These cameras share a wide overlapping field of view, that will not be considered in our tests. Indeed, no inter-image correspondences are taken into account as it is the case in the tests with our hybrid vision system.

The metric used to quantify the drift of our approach are the same as implemented on the KITTI development kit. The rotational error is computed in the following fashion:

$$\varepsilon_R = acos\left(\frac{1}{2}\left(tr(\boldsymbol{R}_R^{-1}\boldsymbol{R}_{GT}) - 1\right)\right), \tag{15}$$

with $\boldsymbol{R}_{GT}$ and $\boldsymbol{R}_R$ the rotations from the ground truth and from our algorithm respectively. While ε_R stands for the rotational error.

[1] http://www.cvlibs.net/datasets/kitti/

The translation error is the euclidean distance (in meters) between the measured position and the real position (from the ground truth):

$$\varepsilon_t = \sqrt{\sum(\boldsymbol{t}_R - \boldsymbol{t}_{GT})^2}. \tag{16}$$

with $\boldsymbol{t}_{GT}$ and $\boldsymbol{t}_R$ the translations from the ground truth and our method respectively. While ε_t is the translation error.

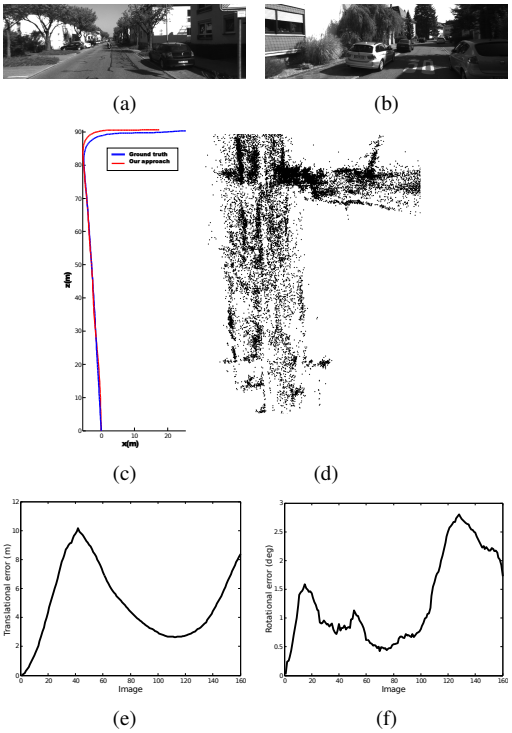
The figure 6 shows results obtained on a sequence of one hundred meters long composed of 160 images. The red line corresponds to the results obtained from our method while the blue line is the ground truth. Note that in this simple sequence we get both a fairly good estimation of the scale but also over the motion of our cameras as emphasized in figure 6. However, we can see a drift over the sequence, this can be corrected by a bundle adjustment refinement unused in this case.

The figure 7 depicts the results obtained with another sequence, where the vehicle travels around 165m during which 204 images were acquired. These results are particularly satisfying, despite a drift in the estimate of the translation and the rotation over time.

These evaluations proved the validity of our approach for the estimation of the motion with real scale without overlapping field of view between cameras. In comparison with the results available on the website KITTI, the developed method is generally less effective than conventional stereo-vision approaches. However, it is much more generic through the combined use of the spherical model -suitable for all SVP cameras- and a non-overlapping SFM method compatible with all calibrated stereo-vision system.

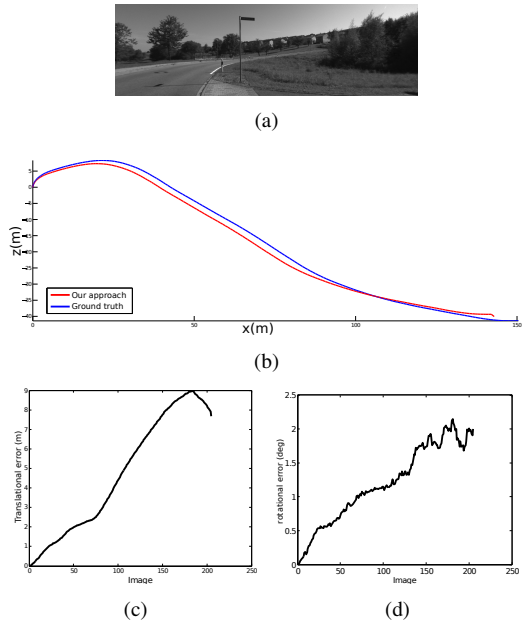
5. Conclusion

In this work we described a novel method for 3D reconstruction and navigation for hybrid vision-system which overcome the problem of stereo correspondence by exploiting the pre-calibration of the rig through a non-overlapping based SFM approach. Furthermore, the proposed method is very versatile since it is suitable for any configuration of cameras and can be easily extended to a larger number of cameras. The experiments with synthetic and real data show the efficiency of the developed algorithm. This work can be extended by taking into consideration the overlapping parts between the two images. Indeed, from our work it is possible to initialise a dense registration approach based on quadri-focal tensor. This additional process may lead to a more accurate estimation of the displacement. It is however essential to choose a metric robust to the strong dissimilarity between omnidirectional and perspective views. For instance, the mutual information which is a particularly well adapted metric for this cameras configuration.



(a) (b) (c) (d) (e) (f)

Fig. 6 Results obtained on a KITTI dataset sequence of 204 images, (a-b) images sample, (c) Estimated trajectory from visual odometry, (d) overview of the 3D reconstruction (top view), (e) Translational error, (f) rotational error



(a) (b) (c) (d)

Fig. 7 Results obtained on a KITTI dataset sequence of 204 images, (a) image from the sequence, (b) Estimated trajectory from visual odometry, (c) Translational error, (d) rotational error

References

[1] J. Gluckman and S. K. Nayar, "Ego-motion and omnidirectional cameras," in *ICCV*, 1998.

[2] D. Eynard, P. Vasseur, C. Demonceaux, and V. Frémont, "Uav altitude estimation by mixed stereoscopic vision," in *IROS*, 2010.

[3] S. Krotosky and M. Trivedi, "On color-, infrared-, and multimodal-stereo approaches to pedestrian detection," *ITS*, 2007.

[4] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox, "Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments," *IJRR*, 2012.

[5] S. Hengstler, D. Prashanth, S. Fong, and H. Aghajan, "Mesheye: a hybrid-resolution smart camera mote for applications in distributed intelligent surveillance," in *IPSN*, 2007.

[6] X. Chen, J. Yang, and A. Waibel, "Calibration of a hybrid camera network," in *ICCV*, 2003.

[7] G. Adorni, L. Bolognini, S. Cagnoni, and M. Mordonini, "Stereo obstacle detection method for a hybrid omni-directional/pin-hole vision system," in *RoboCup 2001*, pp. 244–250, 2002.

[8] F. Roberti, J. Toibero, C. Soria, R. Vassallo, and R. Carelli, "Hybrid collaborative stereo vision system for mobile robots formation," *IJARS*, 2010.

[9] G. Klein and D. Murray, "Parallel tracking and mapping for small ar workspaces," in *ISMAR*, 2007.

[10] A. Geiger, J. Ziegler, and C. Stiller, "Stereoscan: Dense 3d reconstruction in real-time," in *IV*, 2011.

[11] B. Kitt, A. Geiger, and H. Lategahn, "Visual odometry based on stereo image sequences with ransac-based outlier rejection scheme," in *IV*, 2010.

[12] A. Comport, E. Malis, and P. Rives, "Accurate quadrifocal tracking for robust 3d visual odometry," in *ICRA*, 2007.

[13] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, 2013.

[14] C. Demonceaux and P. Vasseur, "Omnidirectional image processing using geodesic metric," in *ICIP*, 2009.

[15] J. Cruz-Mota, I. Bogdanova, B. Paquier, M. Bierlaire, and J. Thiran, "Scale invariant feature transform on the sphere: Theory and applications," *IJCV*, 2012.

[16] B. Clipp, J. Kim, J.-M. Frahm, M. Pollefeys, and R. Hartley, "Robust 6dof motion estimation for non-overlapping, multi-camera systems," in *WACV*, 2008.

[17] R. Hartley and A. Zisserman, *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[18] D. Nistér and F. Schaffalitzky, "Four points in two or three calibrated views: Theory and practice," *IJCV*, 2006.