Hybrid Focal Stereo Networks for Pattern Analysis in Homogeneous Scenes

Emanuel Aldea, K. H. Kiyani

▶ **To cite this version:**

Emanuel Aldea, K. H. Kiyani. Hybrid Focal Stereo Networks for Pattern Analysis in Homogeneous Scenes. Computer Vision - ACCV 2014 Workshops, Nov 2014, Singapour, Singapore. hal-01691982

HAL Id: hal-01691982
https://hal.science/hal-01691982

Submitted on 24 Jan 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Hybrid Focal Stereo Networks for Pattern Analysis in Homogeneous Scenes

Emanuel Aldea[1,3] and Khurom H. Kiyani[2,3]

[1] Autonomous Systems Group, Université Paris Sud, France
[2] Communications and Signal Processing Group, Imperial College London, UK
[3] AquaMed Research and Education, Doha, Qatar

Abstract. In this paper we address the problem of multiple camera calibration in the presence of a homogeneous scene, and without the possibility of employing calibration object based methods. The proposed solution exploits salient features present in a larger field of view, but instead of employing active vision we replace the cameras with stereo rigs featuring a long focal analysis camera, as well as a short focal registration camera. Thus, we are able to propose an accurate solution which does not require intrinsic variation models as in the case of zooming cameras. Moreover, the availability of the two views simultaneously in each rig allows for pose re-estimation between rigs as often as necessary. The algorithm has been successfully validated in an indoor setting, as well as on a difficult scene featuring a highly dense pilgrim crowd in Makkah.

1 Introduction

The problem of multiple camera calibration has been a central topic for the pattern recognition and robotics communities since their inception. Moreover, the use of camera networks has become pervasive in our society; beside their use in surveillance and security enforcement, cameras are heavily relied upon in application domains related to entertainment and sports, geriatrics and elderly care, the study of natural and social phenomena, etc. Motivated by all these developments, a large body of work has been devoted to the problem of estimating accurately the camera network topology, i.e. camera positions and orientations in a common reference system. Inferring the topology in camera networks with non-overlapping fields of view (FOV) is a topic specific to wide-area tracking relying more on high-level image processing and statistical inference and will not be addressed in the current work; the focus of the current article is on estimating the geometric topology for cameras with overlapping FOV. Although such a network may be composed of a large number of cameras firmly attached to a mobile object such as a robot, car, or UAV, most commonly camera networks are static and point towards a specific scene of interest. In these cases, multiple camera calibration is performed by using a specific calibration pattern or object [1–3], which is deployed and moved in the scene during a dedicated calibration phase. If the use of a calibration object is not possible, scene based calibration

may be performed by exploiting visible interest points in methods based on pose refinement [4], or if applicable by using dynamic silhouettes, such as in [5].

A homogeneous scene is defined as an environment lacking completely salient features which would allow their association in different camera views, and which would thus allow for scene based calibration. Typical examples are liquid flow, vapour flow, plant canopy (crops, jungle) or high-density crowds, the latter being the context we have chosen for illustrating our work. The analysis of a homogeneous scene which is not directly accessible for setting up markers, or where the use of calibration objects is not feasible, raises a problem which is not solved by the common methods employed for multiple camera calibration. We approach this problem by replacing cameras with hybrid static stereo rigs, where a long focal camera is used for analysis and a large FOV camera is used for registration with other rigs. By proposing this solution, we avoid using active cameras which require complex models for the dynamic evolution of their intrinsic parameters. Other benefits of possessing simultaneous large and small FOV images of the scene are the fact that the registration does not assume anything about the analysed scene, the fact that the salient features do not have to be static as long as the cameras are accurately synchronized, but then if they are static they can be used to re-estimate continuously the pose and correct phenomena such as camera shaking.

The outline of the paper is as follows. In Section 2 we illustrate the fundamental problem that we address, and discuss related work and alternative solutions. Then, Section 3 recalls the fundamental notions which are required for scene based calibration and for the understanding of the proposed algorithm, which is presented in Section 4. Section 5 illustrates a small scale experiment as well as an application of the proposed algorithm to the analysis of a highly crowded scene, and Section 6 presents the conclusions.

2 Motivation and related work

Based on the simple pinhole projection model (also recalled in Section 3), let us illustrate an issue related to the representation of a homogeneous region of interest in a camera sensor (for all the following tests and examples we will employ Sony ICX274 sensors with a 8.923 mm diagonal and an effective pixel resolution of 1624×1234). We have acquired from the same position and with the same camera three shots using lenses with 4, 8 and 12 mm focals respectively. In the left column of Fig. 1 we present from top to bottom three 50×50 pixel patches from the shots taken with increasing focal lengths. The adjacent images from left to right show areas from these patches (of initial size 20×30 and zoomed for visualization purposes with no interpolation applied). In this case, the long focal lenses are required for retrieving with enough detail entities such as body parts, bags etc. which are essential for a wide range of tasks related to action understanding, monitoring, tracking and surveillance.

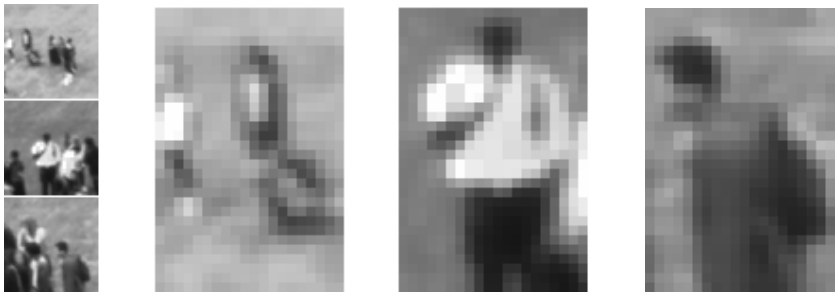


Fig. 1. Left column, from top to bottom: 50×50 pixel patches from shots taken with increasing focal lengths ($f = 4mm$, $f = 8mm$ and $f = 12mm$) from the same position. The following images, from left to right: interest areas from the three previous patches, of initial size 20×30 and zoomed for visualization purposes with no interpolation applied. Detailed features essential for scene analysis are not retrieved below a certain focal length.

From the above illustrations, we note that a wide FOV is beneficial for accurate registration in a camera network, whilst a narrow FOV is beneficial for retrieving the details from the area of interest. By lacking salient features the narrow FOV is not able to estimate robustly or at all the relative pose between multiple cameras.

A calibration pattern visible from all views set on the area of interest can solve the relative pose problem. However, there are multiple applications where this solution is not practical. The area of interest may be far and thus quite large, or it may be inaccessible. During the analysis, the camera poses might change accidentally due to shocks, periodically due to vibrations, or by design (mobile observers); all these scenarios require frequent relative pose estimation updates. In the following paragraphs, we recall briefly some works that are relevant for the problem of multiple view detailed analysis *and* relative pose estimation, highlighting their respective benefits and shortcomings for this scenario.

Zooming cameras One possible solution is to deploy a network of cameras which use motorized zoom lenses. One major consequence of the zooming process is the variation of intrinsic and distortion parameters of the cameras, which have to be re-estimated. Various solutions for zooming recalibration based on the scene have been proposed [6–12]; these solutions are often denoted as self-calibration methods. Beside the fact that all these methods make simplifying assumptions about a subset of the varying parameters, the fundamental limitation is that they still require the continuous presence of salient features, usually interest points or straight lines, in order to operate.

Pan-Tilt-Zoom (PTZ) cameras PTZ cameras have a built-in zooming function and are specifically designed for live monitoring. However, tasks such as surveillance or auto tracking do not require necessarily accurate self-calibration. In the

area of PTZ camera network calibration, scene based solutions have also been proposed [13, 14] but they have the same fundamental limitation i.e. requiring the presence of distinctive landmarks in the field of view.

The strategies recalled up to this point propose interesting solutions for self-calibration and camera registration in the presence of a sufficient number of salient features, but they are not applicable for a camera view if the lens is zoomed on a *homogeneous* and/or *dynamic* scene. Although these scenarios are less common, examples of possible applications abound in the study of crowds and of different types of flows encountered in natural phenomena. The underlying idea for the solution we propose is about transferring pose information in a scene-independent manner to the zoomed camera from a secondary camera able to infer its pose. This leads to a straightforward minimal solution based on a rigid stereo rig featuring two cameras, one with a small FOV used for analysis, and one with a large FOV used for registration within a network of such rigs.

Surprisingly, this solution has not been applied to the analysis of homogeneous scenes. Even considering a broader range of applications, the use of hybrid stereo systems featuring large and narrow FOV is limited. We recall here the setup deployed by the STEREO solar observation mission [15, 16], where the hybrid imagers are nonetheless registered accurately using a star catalogue i.e. am inertial frame of reference. In robotics [17] employ a fisheye and a perspective camera on a UAV. More recently, in [18] the hybrid stereo strategy is employed in order to estimate accurately the pose of a moving binocular in order to insert virtual objects realistically. The fundamental difference is that the design of the system proposed in [18] is tailored for minor pose variations (see the use of the IMU and the small error assumptions), which determine small perspective changes in the appearance of the distant scene, thus greatly simplifying the visual odometry. In contrast, our work addresses a problem where large changes in perspective do not allow for such convenient associations in the large FOV cameras (and for any association at all in the narrow FOV cameras).

With respect to the previous works, the solution based on a hybrid stereo system has clear benefits for the analysis of dynamic homogeneous scenes. We do not have to adopt any simplifying assumptions about the variations of intrinsic parameters, and the calibration precision will be maintained at the optimal level provided by state of the art calibration algorithms. Secondly, the extrinsic parameters of each stereo rig can be estimated independently of the scene. In contrast to the scenario of a zooming camera, the availability at each instant of an accurately registered pair of images allows for accurate pose re-estimation between rigs as often as necessary, overcoming the effect of movement and vibrations.

3 Background on scene based pose estimation

The projection model In the following, we will briefly recall the pinhole camera and optical distortion models that we employ. A point in 3D space $\mathbf{X} = [\mathrm{X\ Y\ Z}]^{\mathrm{T}}$

projects within the image space into a pixel $\mathbf{x} = [x\ y]^{\mathrm{T}}$ according to:

$$\begin{pmatrix} \mathbf{x} \\ 1 \end{pmatrix} = \lambda \mathbf{K} \Big[\mathbf{R} \mid -\mathbf{R}\mathbf{C} \Big] \begin{pmatrix} \mathbf{X} \\ 1 \end{pmatrix} \tag{1}$$

with λ being an undetermined scale factor, $\mathbf{R}$ the orientation of the camera and $\mathbf{C}$ the location of its optical center in world coordinates (we also note $\mathbf{t} = -\mathbf{R}\mathbf{C}$), and $\mathbf{K}$ the intrinsic parameters:

$$\mathbf{K} = \begin{bmatrix} f_x & s & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

Above, f_x and f_y are the focal lengths, $[c_x\ c_y]^{\mathrm{T}}$ represents the principal point, and the skew parameter s is considered 0.

In order to switch to different coordinate frames, we rely on elements of SE(3), the group of rigid body transformations in $\mathbb{R}^3$. A transformation matrix $\mathbf{E}$ takes the form:

$$\mathbf{E} = \begin{bmatrix} \mathbf{R}\ \mathbf{t} \\ 0\ 1 \end{bmatrix} \tag{3}$$

Element multiplication amounts to transitive chaining coordinate frame transformations: $\mathbf{E}^{CA} = \mathbf{E}^{CB}\mathbf{E}^{BA}$ would transfer a 3D point in homogeneous coordinates from reference system A to reference system C.

In order to account for radial distortion, the extension of the pinhole model assumes that if the 3D point $\mathbf{X}$ is projected to $[\tilde{x}\ \tilde{y}\ 1]^{\mathrm{T}}$ under the initial assumptions, then $\mathbf{X}$ would be actually imaged to the distorted location $[x_d\ y_d]^{\mathrm{T}}$:

$$\begin{pmatrix} x_d \\ y_d \end{pmatrix} = (1 + \sum_{i=1}^{3} \kappa_i \tilde{r}^{2i}) \begin{pmatrix} \tilde{x} \\ \tilde{y} \end{pmatrix} \tag{4}$$

where $\tilde{r} = \left(\tilde{x}^2 + \tilde{y}^2\right)^{1/2}$. Thus, $(f_x, f_y, c_x, c_y, \kappa_1, \kappa_2, \kappa_3)$ is in most scenarios the suitable parameter set for a full intrinsic calibration.

Epipolar geometry One tool that we will employ in the following sections is the epipolar constraint, which is a direct implication of the projective geometry between two views. It is worth noting that this constraint is independent of the scene structure, depending exclusively on the intrinsic parameters and the relative pose - as long as the the salient features of the scene are static, or as long as the cameras are accurately synchronized.

Considering two projections $\mathbf{x}_1$ and $\mathbf{x}_2$ of the same point $\mathbf{X}$ in cameras C_1 and C_2, the epipolar constraint defines the relationship between the projections as $\mathbf{x}_2^{\mathrm{T}}\mathbf{F}\mathbf{x}_1 = 0$. $\mathbf{F}$ is known as the fundamental matrix [19], which depends explicitly on the calibration parameters in the following way: $\mathbf{F} = \mathbf{K}_2^{-\mathrm{T}}\mathbf{t}_{\times}\mathbf{R}\mathbf{K}_2^{-1}$ where $\mathbf{t}_{\times}$ is the skew-symmetric matrix associated to $\mathbf{t}$.

The main interest of the epipolar constraint is that it does not make any assumptions about the 3D structure of the scene. Thus, compared to other optimisation algorithms that are commonly employed to estimate the relative pose, the determination of $(\mathbf{R}, \mathbf{t})$ using the epipolar geometry provides a practical minimal parametrization and does not require an initialization. However, the result may be used for the initialization of more complex optimisations, such as the bundle adjustment procedure, briefly recalled in the following paragraph.

Bundle adjustment (BA) Assuming a zero-mean Gaussian distribution of the corner detection errors, bundle adjustment [4, 20] is the Maximum Likelihood Estimator for the joint estimation problem of relative camera poses and of observed 3D point locations. The BA procedure will minimize the following reprojection error:

$$\min_{\hat{P}^i, \hat{\mathbf{X}}_j} \sum_{i,j} d\left(\hat{P}^i(\hat{\mathbf{X}}_j), \mathbf{x}_j^i\right)^2 \tag{5}$$

In the error function above, $\hat{\mathbf{X}}_j$ is the location hypothesis for a point observed by the i^{th} camera. The projection function $\hat{P}^i$ related to the pinhole model (accounting for radial distortion too) depends on the i^{th} camera pose; we consider that the intrinsic parameters are known and are not part of the optimization problem. BA will thus minimize jointly for all the possible camera-point pairs (i, j) the distance between the reprojection $\hat{P}^i(\hat{\mathbf{X}}_j)$ and the actual measurement $\mathbf{x}_j^i$. Solving this optimization problem is studied in depth in the literature, and it generally boils down to exploiting the sparsity of its Hessian matrix and to employing an adapted LS algorithm such as Levenberg-Marquardt [21].

Although BA seems like an ideal solution for multiple view pose estimation, it does have some well-known shortcomings that we will briefly discuss in connection with our specific aim. One common criticism is related to the computational requirements, but this issue is more prevalent in large scale robotics applications, especially if there are real-time constraints to take into account. For a relatively small camera network, the size of the problem is reasonable even for frequent updates. Another important aspect is related to the initialization, which has to be relatively accurate in order to allow the problem to converge to the correct solution. In order to cope with this, we will rely on an initialization based on the epipolar constraint discussed above, but other options are possible too (see for example [22], or [23] if 3D information about some scene features is available). Finally, some practical aspects are equally relevant. Given the high number of parameters which are usually involved, constraining the relative pose variables is more effective if the adjacent camera views for the large FOV cameras are close enough in order to allow for a significant FOV overlap. Stability is also improved if the corner correspondences are spread onto the common field of view.

4 The proposed algorithm

Outline Let us consider a network of N hybrid stereo rigs, the i^{th} rig featuring a small FOV camera C_i^s used for analysis, and a large FOV camera C_i^l employed

for pose estimation in a global frame. The aim of the following procedure is to align accurately the cameras $\{C_1^s, C_2^s, \ldots, C_N^s\}$. We assume that for each rig, the cameras C_i^s and C_i^l have been calibrated. In the following, we will denote by $\mathbf{E}_i^{sl}$ the transform that transfers a point from the large FOV camera to the analysis camera on the i^{th} stereo rig. Also, $\mathbf{E}_{ji}^l$ and $\mathbf{E}_{ji}^s$ are transforms that transfer points from the i^{th} rig to the j^{th} between the large FOV cameras, and respectively between the analysis cameras.

The fact that the stereo rigs are passive allows for a precise intrinsic and extrinsic calibration which can be performed independently of the scene in a controlled environment. Thus the intrinsic parameters $\mathbf{K}_i^s$, $\mathbf{K}_i^l$ as well as the rigid transform $\mathbf{E}_i^{sl}$ that projects a 3D point from the pose estimation camera of the rig to the analysis camera are considered as known.

For the next step, let us consider a pair of spatially adjacent rigs (i, j); in most scenarios, cameras are spread as much as possible, and thus it is necessary to consider adjacent pairs in order to obtain enough reliable interest point matches. Due to initialization requirements, we cannot apply BA directly in order to estimate $\mathbf{E}_{ji}^l$ between the two large FOV cameras on the rigs. We perform SIFT detection and matching [24], and use the normalized 8-point algorithm [25] with RANSAC [26] for robustness to outliers. For the matching step, we employ two filtering strategies based on the uniqueness assumption (the ratio τ of the similarity scores for the top two candidates [24]) and on married matching (both features are the top candidate for each other [27]). Then, we decompose the fundamental matrix [20, chap. 9] and choose the correct solution based on the chirality constraint [28]. Let us denote $\tilde{\mathbf{E}}_{ji}^l$ the rigid transformation estimated after this step. Using $\tilde{\mathbf{E}}_{ji}^l$, and based on the inlier set of matches that were validated during the RANSAC procedure, we build a set of 3D points $\tilde{\mathbf{X}}_{ji}$ by linear triangulation [29].

At this point, we can employ BA using $\tilde{\mathbf{E}}_{ji}^l$ and $\tilde{\mathbf{X}}_{ji}$ as initial estimates, and we obtain a refined relative pose estimation $\hat{\mathbf{E}}_{ji}^l$ for the large FOV cameras in the pair of rigs (i, j). Ideally, BA involving more than a pair of rigs should be performed afterwards whenever possible; however, in a typical setting, cameras are spaced as much as possible around a scene, the limit being imposed by common FOV considerations and the performance of the interest point matching procedure. Thus, we may assume that in most situations non-adjacent rigs will have difficulties for the matching procedure, and will have matches corresponding to disjoint sets of 3D points, which effectively yields the BA problems independent. A particular setting is that of a scene surrounded in a full circle by rigs, and in this case a full BA may be beneficial.

Having the BA estimations, it is trivial to express the C_i^l poses in a common reference system; in the following, we set this reference system as depicted by the position and orientation of C_1^l. Let $\hat{\mathbf{E}}_i^l$ be the rigid transform that links C_1^l to C_i^l. For any two rigs (i, j), we can now use the extrinsic calibrations $\mathbf{E}_i^{sl}$, $\mathbf{E}_j^{sl}$ and the global allignment of the large FOV cameras in order to infer the global allignment of the analysis cameras in the same reference system, as well as their

relative pose:

$$\mathbf{E}_i^s = \mathbf{E}_i^{sl}\hat{\mathbf{E}}_i^l; \mathbf{E}_j^s = \mathbf{E}_j^{sl}\hat{\mathbf{E}}_j^l; \tag{6}$$

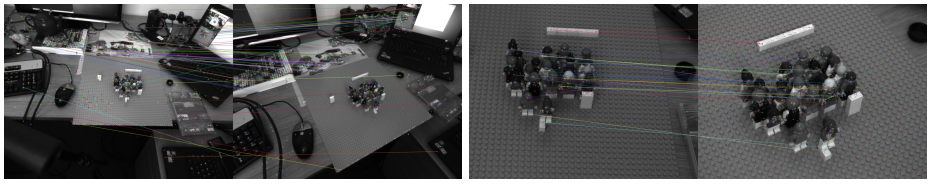
$$\mathbf{E}_{ji}^s = \mathbf{E}_j^{sl}\hat{\mathbf{E}}_j^l \left(\mathbf{E}_i^{sl}\hat{\mathbf{E}}_i^l\right)^{-1} \tag{7}$$

Enforcing a common scale BA can estimate accurately the relative pose up to an unknown scale factor. This limitation applies to the $\hat{\mathbf{E}}_{ji}^l$ estimates only; the values $\mathbf{E}_i^{sl}$ that specify the baseline for cameras on the same rig are not concerned as long as a known size calibration pattern is used for stereo extrinsic calibration. Since the different BA procedures depicted in the following paragraphs are typically independent, we have to enforce a common scale factor among all optimizations using additional information. Depending on the application, it is easier to adopt one of the following strategies. For a small sized scene, we may add a known size object in the common FOV of C_i^l and C_j^l; we thus use $\tilde{\mathbf{X}}_{ji}$ in order to impose a metric scale to the reconstruction. For a large scene, we may either use a similar approach as for the previous setting, or if it is not applicable we may measure the distance between C_i^l and C_j^l (using for example a laser rangefinder), thus using $\tilde{\mathbf{t}}_{ji}$ in order to impose a metric scale to the reconstruction.

5 Experimental results

A small scale scenario We have created a simple example in an indoor environment, using LEGO figurines placed closely in the middle of a homogeneous surface. We have used two hybrid stereo rigs and taken a snapshot of the figurines and surrounding environment. The resulting images are presented in Figure 2: the upper and lower rows show the views from the large FOV (C_1^l,C_2^l) and small FOV (C_1^s,C_2^s) cameras respectively. We have also highlighted the results of the matching procedures; the first matching set ($\tau = 0.4$ for uniqueness) is required for the matching step of the algorithm, while the second set (a more permissive value $\tau = 0.75$ has been used in order to have enough matches) is not used in the algorithm - as the scene is supposed to be poor in salient features - but it is used *exclusively as ground truth for validating the result of the algorithm.* We apply the steps highlighted in the previous Section in order to compute $\mathbf{E}_{21}^s$: estimation of $\mathbf{E}_{21}^l$ using SIFT matching followed by decomposition of $\mathbf{F}_{21}^l$ and BA, then exploitation of $\mathbf{E}_1^{sl}$ and $\mathbf{E}_2^{sl}$ provided by stereo calibration, and also the setup of the right scale by using an object of known size (the long brick of length 79.8 mm).

In order to estimate numerically the quality of the rigid transform $\mathbf{E}_{21}^s$ obtained, we have exploited the matches that we were able to determine directly between the small FOV cameras in this example. In homogeneous scenes, interest points may be completely absent, or the scarcity of matches may have a detrimental effect on the stability of the estimation of $\mathbf{F}$. Therefore, in our example we set up a base BA problem between C_1^s and C_2^s where we initialize the system by the decomposition of $\mathbf{F}_{21}^s$. Alternatively, we use the rigid transform $\mathbf{E}_{21}^s$ as initialization for the triangulation of matches and for the BA procedure.



a) b)

Fig. 2. A set of images used for pose estimation in a simple indoor environment; the images in a) correspond to C_1^l and C_2^l, and the images in b) show the images captured by C_1^s and C_2^s. Both pairs of images have been matched using SIFT; the first set of matches are necessary for the algorithm, whilst the second set is used *exclusively as ground truth for validating the result of the algorithm.*

The resulting solutions and mean reprojection errors for these two scenarios are presented in Table 1.

As we notice, the two optimization problems converge towards the same solution, but $\mathbf{E}_{21}^s$ brings the optimization much closer to the objective in terms of mean reprojection error. This result is interesting for a number of reasons. Firstly, even though we do not have a case of optimization stuck in a local minimum due to the worse initialization, this is a good example of coarse to fine resolution of the relative pose estimation. This approach is helpful for robotics applications in case of unstable optimizations (few matches in the small FOV cameras), and also interesting for the computation gain due to a faster convergence of BA (25 iterations with initial subpixel mean reprojection error compared to 37 iterations). Secondly, and most importantly, this example shows that in cases where we can not compute $\mathbf{E}_{21}^s$ directly due to the complete absence of salient features, we are able using this algorithm to infer the unknown rigid transform from the adjacent large FOV cameras with a high level of accuracy.

Table 1. Relative poses between C_1^s and C_2^s. The Euler angles are expressed in degrees, and the mean reprojection errors in pixels. Tilde values represent estimations prior to the BA procedure, and hat values denote estimations refined by BA. The difference between the two rows consists in the initialization of BA; in the first case we use the SIFT matches depicted in Figure 2b), whilst in the second case we use the result of our algorithm.

$(\tilde{\psi};\tilde{\boldsymbol{\theta}};\tilde{\phi})$	$\tilde{\mathbf{C}}$	$\tilde{\epsilon}$	$(\hat{\psi};\hat{\boldsymbol{\theta}};\hat{\phi})$	$\hat{\mathbf{C}}$	$\hat{\epsilon}$	**Iter.**	**Observations**
$\begin{pmatrix}24.13\\21.04\\10.67\end{pmatrix}$	$\begin{pmatrix}-0.89\\-0.30\\0.33\end{pmatrix}$	37.16	$\begin{pmatrix}23.95\\21.13\\3.74\end{pmatrix}$	$\begin{pmatrix}-0.79\\-0.25\\0.55\end{pmatrix}$	0.199	37	Base solution
$\begin{pmatrix}23.85\\16.42\\3.53\end{pmatrix}$	$\begin{pmatrix}-0.77\\-0.23\\0.59\end{pmatrix}$	0.489	$\begin{pmatrix}23.95\\21.13\\3.74\end{pmatrix}$	$\begin{pmatrix}-0.79\\-0.25\\0.55\end{pmatrix}$	0.199	25	Init. by $\mathbf{E}_{21}^s$

Pose estimation for high-density crowds We have deployed two hybrid stereo rigs at the grand mosque in Makkah during very congested times of the Hajj period, in October 2012. The access constraints to the site impose a large perspective change between the two points of observation. As a result, neither SIFT nor even ASIFT [30] algorithms were capable to provide any correct matches which are required as inputs for the algorithm we propose. Consequently, we had to rely on manual matching of salient structures in order to bootstrap the algorithm.

In Figure 3 we present the data our algorithm processed and registered; images in a) and b) correspond to C_1^l and C_1^s, and c) and d) correspond to C_2^l and C_2^s respectively. The large FOV cameras contain enough common salient features, although the perspective variation does not allow for automated matching, and human intervention is necessary. Figure 3e) presents such a user specified correspondence; in total we have used 34 user specified correspondences, of which 26 have been considered inliers for the fundamental matrix evaluation. For visualization purposes, Figures 3f) and 3g) present the central structure with the manually matched features, and the 3D structure of the scene with the camera axis aligned and an approximate representation of the ground plane.

The numerical results of the algorithm for this setting are presented in Table 2; the relative rotations are expressed in degrees using Euler angles, the relative center position is expressed as a unit $\mathbb{R}^3$ vector, and mean reprojection errors are expressed in pixels. Also, as specified in the algorithm outline (Section 4), tilde values represent estimations prior to the BA procedure, and hat values denote estimations refined by BA. The first row corresponds to Step (iii) of the algorithm, the pose estimation between C_1^l and C_2^l. The output values are consistent with the actual location of the cameras; the large angle displacements emphasize the difficulty of the task, and explain as well the limitation of the automated matching procedure in this case.

We have also refined the relative positions of the cameras within the individual rigs. These values are provided by the stereo calibration procedure, and we validated them by performing SIFT matching between the large and small FOV cameras, and by using the stereo calibration pose as an initializer for BA (rows 2 and 3 in Table 2). The threshold for uniqueness filtering has been set as $\tau = 0.3$. However, the stereo calibration performed on site could not be done in optimal conditions. As an alternative solution, we used as pose initializations values that we obtained in the same way as for the first row of Table 2, by estimating and decomposing the fundamental matrix. The solutions obtained are presented on rows 4 and 5 in Table 2. These solutions were more accurate, and finally they have the advantage of requiring only the intrinsic camera parameters. It is worth noting that the stereo baseline is approximately 6 cm, while the distance to the scene is three orders of magnitude higher, and in these circumstances the relative angles and not the camera center relative positions will be the most relevant for scene based estimation.

Having thus obtained all the necessary relative poses (rows 1, 4 and 5 in Table 2), we are able to estimate the relative pose between the long focal cameras. Ground truth estimations are not possible, but in order to estimate the accuracy

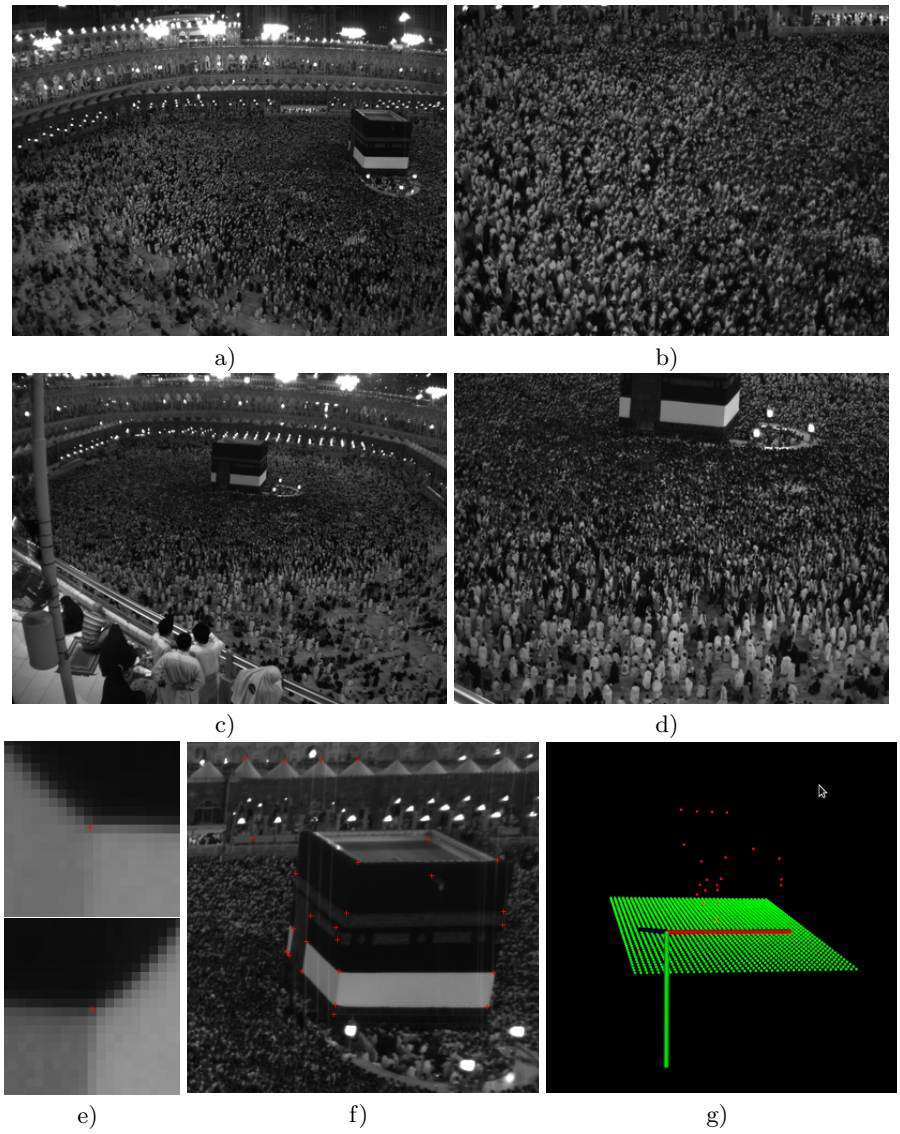


Fig. 3. A set of images used for pose estimation; the images in a) and b) correspond to C_1^l and C_1^s, and c) and d) correspond to C_2^l and C_2^s respectively. An example of user specified correspondences is illustrated in e). In f) we present the interest points used in the central region of one of the images, and in g) the inferred camera orientation (RGB axis for XYZ), with the approximate ground plane highlighted in green, for easier visualization.

Table 2. Relative poses between analysis cameras placed on different rigs (first row), and between cameras placed on the same rig (rows 2-5). The Euler angles are expressed in degrees, and the mean reprojection errors in pixels. Tilde values represent estimations prior to the BA procedure, and hat values denote estimations refined by BA. The difference between the rows 2-3 and 4-5 consists in the initialization of the BA; in the first case we use the stereo calibration, whilst in the second case we use directly the images, in the same way as for the first row initialization.

Cam. pair	$(\tilde{\boldsymbol{\psi}};\tilde{\boldsymbol{\theta}};\tilde{\boldsymbol{\phi}})$	$\tilde{\mathbf{C}}$	$\tilde{\boldsymbol{\epsilon}}$	$(\hat{\boldsymbol{\psi}};\hat{\boldsymbol{\theta}};\hat{\boldsymbol{\phi}})$	$\hat{\mathbf{C}}$	$\hat{\boldsymbol{\epsilon}}$	**Observations**
$C_1^l \Rightarrow C_2^l$	$\begin{pmatrix}53.27\\71.52\\32.94\end{pmatrix}$	$\begin{pmatrix}-0.78\\-0.19\\0.59\end{pmatrix}$	4.00	$\begin{pmatrix}59.68\\69.11\\42.75\end{pmatrix}$	$\begin{pmatrix}-0.81\\-0.18\\0.55\end{pmatrix}$	0.25	Manual Init.
$C_1^l \Rightarrow C_1^s$	$\begin{pmatrix}-0.37\\-0.58\\0.51\end{pmatrix}$	$\begin{pmatrix}0.79\\0.00\\-0.62\end{pmatrix}$	1.017	$\begin{pmatrix}-0.33\\-0.34\\0.48\end{pmatrix}$	$\begin{pmatrix}0.11\\-0.01\\-0.99\end{pmatrix}$	0.076	Stereo Calib. Init.
$C_2^l \Rightarrow C_2^s$	$\begin{pmatrix}-0.19\\0.72\\0.23\end{pmatrix}$	$\begin{pmatrix}0.94\\0.05\\-0.34\end{pmatrix}$	1.661	$\begin{pmatrix}-0.09\\0.71\\0.12\end{pmatrix}$	$\begin{pmatrix}0.57\\0.23\\0.78\end{pmatrix}$	0.252	Stereo Calib. Init.
$C_1^l \Rightarrow C_1^s$	$\begin{pmatrix}-0.36\\-0.23\\0.43\end{pmatrix}$	$\begin{pmatrix}0.01\\-0.06\\0.99\end{pmatrix}$	0.096	$\begin{pmatrix}-0.33\\-0.28\\0.47\end{pmatrix}$	$\begin{pmatrix}0.06\\-0.02\\-0.99\end{pmatrix}$	0.084	SIFT Matching
$C_2^l \Rightarrow C_2^s$	$\begin{pmatrix}-0.16\\0.74\\0.10\end{pmatrix}$	$\begin{pmatrix}-0.02\\-0.01\\-0.99\end{pmatrix}$	0.097	$\begin{pmatrix}-0.13\\0.75\\0.10\end{pmatrix}$	$\begin{pmatrix}-0.01\\0.00\\-0.99\end{pmatrix}$	0.087	SIFT Matching

of the result we have located in the crowd a number of salient elements (either distinctive heads, or distinctive configurations of people) and we illustrate the result by drawing for each feature the epipolar line, and judging by its proximity to the corresponding feature in the other image. In Figure 4, the upper row corresponds to elements identified in C_1^s (Figure 3b) and the lower row presents the same elements identified in C_2^s (Figure 3d). The following remarks are necessary at this point. Firstly, the perspective change makes the correspondence search very tedious even for a human. Secondly, the drawing of the epipolar line has actually assisted us in pinpointing most of these correspondences, and we are confident that the method will be helpful in automating these tasks.

Discussion of the dense crowd results Overall, the distance in the image space between the corresponding element and the corresponding epipolar line is in the range of a few pixels. The major factors responsible for these misalignments are the inaccuracies in estimating the intrinsic parameters, as well as the errors related to the relative pose estimations - but for the dimensions of the scene involved in the experiment, we argue that the results are very promising.

Moreover, some areas of the scene exhibit near perfect alignments. The first four matches presented in Figure 4 (the white cap man in a) positioned under the epiline, in the left part of the patch; the person in b) looking slightly towards the left; the woman in c) wearing a white veil, and positioned in front of two other women similarly dressed; the woman in d) wearing a white veil, and positioned with the back towards the second camera) are very accurate, in spite of the fact that in one of the images the first three persons are located near the border, a fact which potentially increases radial distortion related errors.

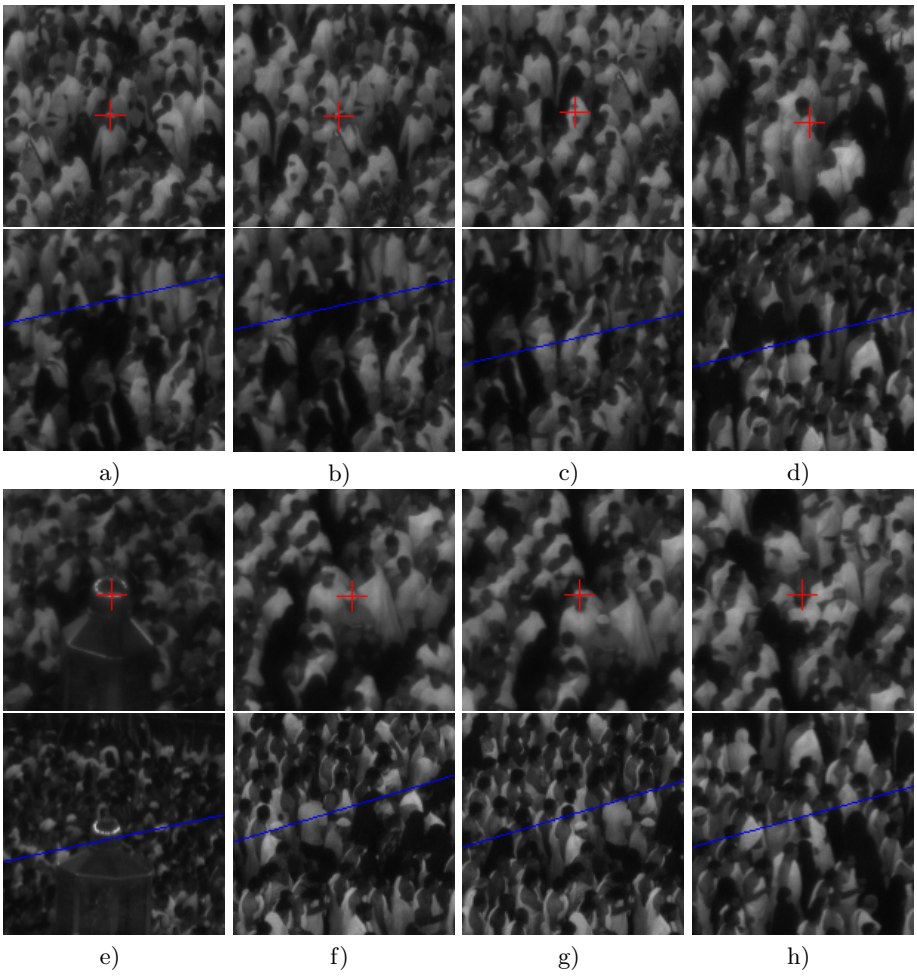


Fig. 4. A number of pixel-epipolar line correspondences between the two analysis cameras presented in Figure 3b) and 3d). Ideally, the correspondent of a point highlighted by the red cross in the upper row should be situated along the blue epipolar line visible in the lower row image. These results are discussed in Section 5.

We could also identify the following correspondences which exhibit small but visible misalignments: the shiny circumference of the Station of Ibrahim, depicted in e); another two men wearing white caps, presented in f) and g); a distinctively bearded man presented in h).

The fact that the epipolar line does not pass precisely through the corresponding element is not detrimental for the purpose of association and tracking in the crowd. Assuming that the person is not occluded, using this extra information we would not only be able to trim down the research space to a band

along the epipolar line, but also if we were able to position the ground plane within the same coordinate system we would further reduce the research space to a fraction of the band. Of course, in order to do dense matching reliably we still need a neighborhood based similarity measure that has to be resilient to major perspective change and occlusions; this is a promising direction of research that we intend to follow in order to benefit from the relative pose algorithm we propose, and ultimately in order to perform dense associations.

Finally, the present results of the proposed algorithm on this type of data are also encouraging as they illustrate the potential of multi-camera systems in extremely crowded environments. In the current research context, this application field has been associated mostly with single camera systems [31], but paradoxically it would greatly benefit from multi-view systems given the frequent occlusions and scene clutter that characterize it.

6 Conclusion

In this paper we propose a new method for aligning multiple cameras analysing a homogeneous scene. Our method addresses the settings where for practical reasons calibration pattern/object based registrations are not possible. By employing stereo rigs featuring a long focal analysis camera and a short focal registration camera, the proposed solution alleviates the requirement to get access to the studied scene. The fact that we are using a large FOV simultaneously allows us to avoid making any assumptions about the homogeneous region we analyse, such as the presence of shades, silhouettes etc. A first experiment has been conducted in an indoor environment and has shown, by using interest point correspondences in the analysis area as ground truth, that this method can guide the relative pose estimation for scenes poor in salient features in a coarse-to-fine manner supported by hardware. The second test has shown that in the absence of any salient features, the method is capable of providing a full calibration of the analysis cameras in a difficult, large scale scenario.

In the future, we would like to investigate the applicability of the proposed hybrid stereo solution in two frequently recurring settings. We intend to employ this method as a preprocessing step for a wide range of homogeneous pattern analysis applications, such as those related to the extraction of accurate models for highly dense crowd dynamics. Secondly, we would like to evaluate further the potential of this solution in specific applications such as autonomous robot navigation or image alignment and stitching, which employ pyramid based coarse-to-fine optimizations; our setup augments these systems by supplementing the image pyramid with a level provided by an independent data source.

Acknowledgement. This work was funded by QNRF under the grant NPRP 09-768-1-114. The authors acknowledge A. Gutub and his team at the Centre of Research Excellence in Hajj and Omrah; and O. Gazzaz, F. Othman, A. Fouda and B. Zafar at the Hajj Research Institute for their organisation and logistical support in the video data collection at the grand mosque in Makkah, as well as for useful discussions.

References

1. Zhang, Z., Zhang, Z.: A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence **22** (1998) 1330–1334
2. Baker, P., Aloimonos, Y.: Complete calibration of a multi-camera network. In: Omnidirectional Vision, 2000. Proceedings. IEEE Workshop on. (2000) 134–141
3. Svoboda, T., Martinec, D., Pajdla, T.: A convenient multi-camera self-calibration for virtual environments. PRESENCE: Teleoperators and Virtual Environments **14** (2005) 407–422
4. Triggs, B., McLauchlan, P.F., Hartley, R.I., Fitzgibbon, A.W.: Bundle adjustment - a modern synthesis. In: Workshop on Vision Algorithms. (1999) 298–372
5. Sinha, S., Pollefeys, M.: Camera network calibration and synchronization fromsilhouettes in archived video. International Journal of Computer Vision **87** (2010) 266–283
6. Sturm, P.F.: Self-calibration of a moving zoom-lens camera by pre-calibration. Image Vision Comput. **15** (1997) 583–589
7. Ahmed, M., Farag, A.: Nonmetric calibration of camera lens distortion: differential methods and robust estimation. Trans. Img. Proc. **14** (2005) 1215–1230
8. Wang, A., Qiu, T., Shao, L.: A simple method of radial distortion correction with centre of distortion estimation. J. Math. Imaging Vis. **35** (2009) 165–172
9. Kukelova, Z., Pajdla, T.: A minimal solution to radial distortion autocalibration. Pattern Analysis and Machine Intelligence, IEEE Transactions on **33** (2011) 2410–2422
10. Lourakis, M.I., Deriche, Rachid, D.: Camera Self-Calibration Using the Singular Value Decomposition of the Fundamental Matrix: From Point Correspondences to 3D Measurements. Technical Report RR-3748, INRIA (1999)
11. Josephson, K., Byrod, M.: Pose estimation with radial distortion and unknown focal length. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. (2009) 2419–2426
12. Dang, T., Hoffmann, C., Stiller, C.: Continuous stereo self-calibration by camera parameter tracking. Trans. Img. Proc. **18** (2009) 1536–1550
13. Sinha, S.N., Pollefeys, M.: Towards calibrating a pan-tilt-zoom camera network. In: OMNIVIS. (2004)
14. Bimbo, A.D., Dini, F., Lisanti, G., Pernici, F.: Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks. Comput. Vis. Image Underst. **114** (2010) 611–623
15. Eyles, C., Harrison, R., Davis, C., Waltham, N., Shaughnessy, B., Mapson-Menard, H., Bewsher, D., Crothers, S., Davies, J., Simnett, G., et al.: The heliospheric imagers onboard the stereo mission. Solar Physics **254** (2009) 387–445
16. Brown, D., Bewsher, D., Eyles, C.: Calibrating the pointing and optical parameters ofthestereo heliospheric imagers. Solar Physics **254** (2009) 185–225
17. Eynard, D., Vasseur, P., Demonceaux, C., Frémont, V.: Real time uav altitude, attitude and motion estimation from hybrid stereovision. Auton. Robots **33** (2012) 157–172
18. Oskiper, T., Sizintsev, M., Branzoi, V., Samarasekera, S., Kumar, R.: Augmented reality binoculars. In: Mixed and Augmented Reality (ISMAR), 2013 IEEE International Symposium on. (2013) 219–228
19. Faugeras, O.D.: What can be seen in three dimensions with an uncalibrated stereo rig. In: ECCV. (1992) 563–578

20. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. Second edn. Cambridge University Press, ISBN: 0521540518 (2004)
21. Lourakis, M.A., Argyros, A.: SBA: A Software Package for Generic Sparse Bundle Adjustment. ACM Trans. Math. Software **36** (2009) 1–30
22. Nistér, D.: An efficient solution to the five-point relative pose problem. IEEE Trans. Pattern Anal. Mach. Intell. **26** (2004) 756–777
23. Kneip, L., Scaramuzza, D., Siegwart, R.: A novel parametrization of the perspective-three-point problem for a direct computation of absolute camera position and orientation. In: CVPR. (2011) 2969–2976
24. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
25. Hartley, R.: In defense of the eight-point algorithm. Pattern Analysis and Machine Intelligence, IEEE Transactions on **19** (1997) 580–593
26. Fischler, M.A., Bolles, R.C.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24** (1981) 381–395
27. Nistér, D., Naroditsky, O., Bergen, J.R.: Visual odometry. In: CVPR (1). (2004) 652–659
28. Hartley, R.I.: Chirality. International Journal of Computer Vision **26** (1998) 41–61
29. Hartley, R.I., Sturm, P.F.: Triangulation. Computer Vision and Image Understanding **68** (1997) 146–157
30. Yu, G., Morel, J.M.: ASIFT: An Algorithm for Fully Affine Invariant Comparison. Image Processing On Line **2011** (2011)
31. Wang, X.: Intelligent multi-camera video surveillance: A review. Pattern Recogn. Lett. **34** (2013) 3–19