Exploiting Single View Geometry in Pan-Tilt-Zoom Camera Networks

A. del Bimbo, F. Dini, A. Grifoni, F. Pernici

▶ **To cite this version:**

A. del Bimbo, F. Dini, A. Grifoni, F. Pernici. Exploiting Single View Geometry in Pan-Tilt-Zoom Camera Networks. Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications - M2SFA2 2008, Andrea Cavallaro and Hamid Aghajan, Oct 2008, Marseille, France. inria-00326775

HAL Id: inria-00326775

https://inria.hal.science/inria-00326775

Submitted on 6 Oct 2008

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.

Exploiting Single View Geometry in Pan-Tilt-Zoom Camera Networks

A. Del Bimbo[1], F. Dini[1], A. Grifoni[2], F. Pernici[1]

[1] MICC University of Florence, Italy
[2] Thales Security Solutions Florence, Italy

Abstract. PTZ (pan-tilt-zoom) camera networks have an important role in surveillance systems. They have the ability to direct the attention to interesting events that occur in the scene. In order to achieve such behavior the cameras in the network use a process known as sensor slaving: one (or more) master camera monitors a wide area and tracks moving targets so as to provide the positional information to one (or more) slave camera. The slave camera foveates at the targets in high resolution. In this paper, we propose a simple method to solve two typical problems that are the basic building blocks to create high level functionality in PTZ camera networks: the computation of the world to image homographies and the computation of camera to camera homographies. The first one is used for computing the image sensor observation model in sequential target tracking, the second one is used for camera slaving. Finally a cooperative tracking approach exploiting the use of both homographies is presented.

1 Introduction

In the last few years, with the advent of smart, computer-enabled surveillance technologies and IP cameras, the use of surveillance systems for security reasons has exploded. Moreover control equipment such as PTZ cameras (also known as dome camera) are and will be of invaluable help for monitoring wide outdoor areas with a minimal number of sensors. For these cameras, however, precalibration is almost impossible. In fact, transportation, installation, changes in temperature and humidity as present in outdoor environments, typically affect the estimated parameters. Moreover, it is impossible to recreate the full range of zoom and focus settings. A tradeoff has to be made for simplicity against strict geometric accuracy.

It is well known that in areas where the terrain is planar the relation between image pixels and terrain locations or between the image pixels of two cameras, is a simple 2D homography. While finding at least four well distributed image point features to compute the world to image mapping is easy in an indoor environment (provided that calibration grids are pasted on the floor of the room), in outdoor environments this is proved to be more complicated especially if the scene area is not sufficiently textured. For the same reason also image to image homographies between two different views taken from the same camera or different cameras are not easy to estimate.

In this paper we propose a calibration and a tracking method for PTZ cameras that greatly simplifies cooperative target tracking. The key contributions

of the paper are threefold: first, we show how to combine single view geometry and planar mosaic geometry to compute the world to image and the image to image homography (The first is used for computing image sensor likelihood for sequential target tracking, the second is used for camera slaving). Second we show how line features in the mosaic computed with a parameterization with a minimum number of parameters gives globally better results than using 3D known coordinate of human-measurable feature points. Third the method can be used to enhance future surveillance systems to keep track of targets at high resolution that may not necessarily be captured within one field of view.

2 Related Work

In the literature, several methods exist to calibrate one or several PTZ cameras. These methods can be distinguished according to the particular task to perform.

The method [8] can be used to self-calibrate (without calibration targets) a single PTZ camera by computing the homographies induced by rotating and zooming the camera. In [6] the same approach has been analyzed considering the effect of imposing different constraints on the intrinsic parameters of the camera. They reported that best results are obtained when the principal point is assumed to be constant throughout the sequence although it is known to be varying in reality. In [12] a very thorough evaluation of the same method is performed with more than one hundred images. Then the internal calibration of the two PTZ camera is used for 3D reconstruction of the scene through essential matrix and triangulation by using the mosaic images as a stereo pair.

Another class of methods using self-calibration based on moving objects has been proposed. For example [5] and [14] use LEDs. As the LED is moved around and visits several points, these positions make up the projection of a virtual object (modeled as 3D point cloud) with unknown position. However the need of synchronization prevent the use of the approaches for IP camera networks.

After the VSAM project [2] new methods have been proposed for calibrating PTZ cameras with simpler and more flexible approaches suitable for outdoor environment. These are mainly targeted to high resolution sensing of objects at a distance, and therefore the zoom usage is of mandatory importance in these methods. Of particular interest are the works [15], [7] and [4] where the PTZ camera scheduling problem is addressed.

3 Problem Formulation

In section 3.1 the PTZ camera network with master-slave configuration is defined in terms of its relative geometry. The section 3.2 describe how to compute this basic geometry using the single view and the planar mosaic geometry. Finally in section 4 the estimated geometry is exploited to cooperatively track a target over an extended area at high resolution.

3.1 PTZ Camera Networks with Master-Slave Configuration

PTZ cameras are particularly effective when configured in a master-slave configuration [15]: the master camera is set to have a global view of the scene so

that it can track objects over extended areas using simple tracking methods with adaptive background subtraction. The slave camera, can then follow the trajectory to generate close-up imagery of the object. Evidently their respective roles can be exchanged. The master-slave configuration can be extended to the case of multiple PTZ cameras. Fig.1 shows the pair-wise relationship between two cameras in this configuration. $\mathtt{H}$ is the world to image homography of the master camera , $\mathtt{H}'$ is the homography relating the image plane of $\mathbf{C}$ with the reference image plane $\mathbf{\Pi}'$ of the slave camera $\mathbf{C}'$ and $\mathtt{H}_k$ is the homography relating the reference image plane $\mathbf{\Pi}'$ with the current image plane of the slave camera. Once $\mathtt{H}_k$ and $\mathtt{H}'$ are known the imaged location $\mathbf{x}_1$ of a moving target $\mathbf{X}_1$ tracked by the stationary camera $\mathbf{C}$ can be transferred to the zoomed view of $\mathbf{C}'$ by:

$$\mathtt{T}_k = \mathtt{H}_k \cdot \mathtt{H}' \tag{1}$$

With this pairwise relationship between cameras the number of possible network configuration can be calculated. Given a set of PTZ cameras $\mathbf{C}_i$ viewing a planar scene, we define $\mathcal{N} = \{\mathbf{C}_i^s\}_{i=1}^M$ a PTZ camera network with the master slave relationship, where M denotes the number of cameras in the network and s defines the state of each camera. At any given time these cameras can be in one of two states $s_i = \{\text{MASTER}, \text{SLAVE}\}$.

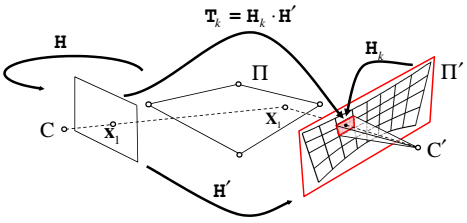


Fig. 1. The pair-wise relationship between two cameras in master-slave configuration. The camera $\mathbf{C}$ is the tracking camera and $\mathbf{C}'$ is the slave camera. $\mathtt{H}$ and $\mathtt{H}'$ are respectively the world to image homography and the image to image homography. $\mathbf{\Pi}$ is the 3D world plane while $\mathbf{\Pi}'$ is the mosaic plane of the slave camera.

The network $\mathcal{N}$ can be in one of $2^M - 2$ possible state configurations. All cameras in a master state, or all cameras in a slave state cannot be defined. It is worth noticing that from this definition more than one camera can act as a MASTER and/or SLAVE camera. In principle without any loss of generality if all the cameras in a network have an overlapping field of view (i.e. they are in a full connected topology) the cameras can be set in a master-slave relationship each other (not only in a one to one relationship). For example in order to cover large areas more master cameras can be placed with adjacent fields of view. In this case if they act as a master camera, one slave camera suffice to observe the whole area. Several master cameras can have overlapping fields of view so as to achieve higher tracking accuracy (multiple observations of the same object from different cameras can taken into account to obtain a more accurate measurement and determine a more accurate foveation by the slave camera). Similarly, more than one camera can act as a slave camera while just one can be used as a master for tracking, for example for capturing high resolution images of moving objects from several viewpoints.

3.2 Minimal PTZ Camera Model Parameterization

We consider the pin-hole camera model projecting the three-dimensional world onto a two-dimensional image, with fixed principal point, without modelling the radial distortion. It is assumed that the camera rotates around its optical center with no translation. The pan and tilt axes are assumed to intersect. The 3×3 matrix $\mathtt{K}_i$ contains the intrinsic parameters of the camera for the image taken at time i and the 3×3 matrix $\mathtt{R}_i$ defines its orientation. It is possible to model the whole projection as $\mathtt{P}_i = [\mathtt{K}_i\mathtt{R}_i \ 0]$, where the equality denotes equality up to a scale factor. As described also in [9] it is possible to derive the inter-image homography, between image i and image j as: $\mathtt{H}_{ji} = \mathtt{K}_j\mathtt{R}_{ji}\mathtt{K}_i^{-1}$. Due to the mechanical nature of PTZ cameras it is possible to assume that there is no rotation around the optical axis: $\theta = 0$. We will assume that the center of projection lies at the image center, the pan-tilt angles between spatially overlapping images is small and the focal length does not vary too much between two overlapping images $f_i = f_j = f$. Under these assumptions, the image-to-image homography can be approximated by:

$$\mathtt{H}_{ji} = \begin{pmatrix} 1 & 0 & f\psi_{ji} \\ 0 & 1 & -f\phi_{ji} \\ \frac{-\psi_{ji}}{f} & \frac{-\phi_{ji}}{f} & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & h_1 \\ 0 & 1 & h_2 \\ h_3 & h_4 & 1 \end{pmatrix} \tag{2}$$

where ψ_{ji} and ϕ_{ji} are respectively the pan and tilt angles from image j to image i, [1]. Each point match contributes with two rows in the measurement matrix. Since there are only four unknowns, $(h_1\ h_2\ h_3\ h_4)$, two point matches suffice to estimate the homography. Estimates for ψ, ϕ and f can be calculated from the entries of $\mathtt{H}_{ji}$. With this parameterization matching and minimization are generally more simple than using the full 8 DOF homography. While with this parameterization calibration parameters are not accurate, it is still possible to create a wide single view (i.e. a planar mosaic) maintaining the projective properties of image formation (i.e. straight lines are still straight lines in the mosaic). This new view, provided that a moderate radial distortion is present, can be considered as a novel wide angle single perspective image.

Recovering The Homographies. It is well known that given three orthogonal vanishing points $\mathbf{v}_1$,$\mathbf{v}_2$,$\mathbf{v}_3$ they can be used to calibrate a natural pinhole camera computing the focal length (1 DOF) and principal point (2 DOFs). This can be done referring to the image of the absolute conic (IAC) ω using the following constraints [10]: $\mathbf{v}_1\omega\mathbf{v}_2 = 0$, $\mathbf{v}_2\omega\mathbf{v}_3 = 0$, $\mathbf{v}_3\omega\mathbf{v}_1 = 0$. Other constraints on ω can be obtained from circles [13] [3] and can be exploited for example in the case of sport video analysis. The ω is responsible for internal camera parameters according to: $\omega = \mathtt{K}^{-T}\mathtt{K}^{-1}$. However as shown below we don't need to compute explicitly the entries of $\mathtt{K}$ for recovering the homographies. When ω and the vanishing line $\mathbf{l}_\infty$ of a 3D world plane are known it is possible to compute up to a similarity transformation the metric structure of the plane. The rectifying homography [10] can be computed from the image of the absolute conic ω and

the vanishing line $\mathbf{l}_\infty$ of the scene plane as:

$$\mathtt{H}_r = \begin{pmatrix} \beta^{-1} & -\alpha\,\beta^{-1} & 0 \\ 0 & 1 & 0 \\ l_1 & l_2 & 1 \end{pmatrix} \quad , \tag{3}$$

where $\mathbf{l}_\infty = (l_1, l_2, 1)$ is the representation of the vanishing line in homogeneous coordinates while α and β are two scalars that can be computed from the imaged circular points $\mathbf{i}$ and $\mathbf{j}$. The imaged circular points are two complex conjugate point pairs (i.e. $\mathbf{i} = \text{conj}(\mathbf{j})$) that are responsible for the metric properties of imaged planes. They are the projection of the circular points $\mathbf{I}$ and $\mathbf{J}$. The circular points $\mathbf{I}$ and $\mathbf{J}$ are in the Euclidean world (the scene plane) at canonical coordinates $\mathbf{I} = (1, i, 0)$, $\mathbf{J} = \text{conj}(\mathbf{I})$. It can be shown that the following relationship holds [10]: $\mathbf{i} = \mathtt{H}_r^{-1}(1, i, 0) = (\alpha - i\beta, 1, -l_2 - l_1\alpha + il_1\beta)$. So the two scalars α and β are directly extracted from the first component of $\mathbf{i}$. The vanishing line $\mathbf{l}_\infty$ is obtained as $\mathbf{l}_\infty = \mathbf{v}_1 \times \mathbf{v}_2$, where $\mathbf{v}_1$ and $\mathbf{v}_2$ are the vanishing points of the two orthogonal directions in the scene plane. The imaged circular points are computed as the intersections of $\mathbf{l}_\infty$ with ω. The transformation of eq.3 relates the world to the image up to an unknown similarity transformation $\mathtt{H}_s$. The $\mathtt{H}_s$ transformation has 4 DOF: two for translation, one for rotation and one for scaling. Two correspondences suffice to compute $\mathtt{H}_s$ (i.e. the world coordinates of two points with their projection onto the mosaic). Without any loss of generality it is possible to choose the first point as the origin $\mathbf{O} = (0, 0)$ and the second point as the distance from the first in the 3D world reference. Operatively, just a length is measured in 3D. The final world to image homography can finally be computed as:

$$\mathtt{H} = \mathtt{H}_r\mathtt{H}_s \tag{4}$$

It is important, for the accuracy of the computation of the vanishing points, to define the reference image where to stitch the mosaic. For example it does not make any sense to choose images where the image plane is either parallel or orthogonal with scene plane or with any scene plane orthogonal to it. In fact in these viewing conditions the vanishing points meet at infinity in the image, producing high uncertainty in their localization.

Fig. 2(b) shows the planar mosaic obtained from a set of images acquired by a PTZ camera (these images are shown in fig.2(a)). In particular in the same figure are also shown the three orthogonal vanishing points $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ and the vanishing line $\mathbf{l}_\infty$ used to obtain the world to image homography $\mathtt{H}$ of eq.4. Fig.2(c) shows the rectified mosaic of the area under surveillance.

For the computation of the inter-image homography, it is necessary to choose four well spaced pairs of corresponding points or lines in the two mosaics. Due to the wide angle view of the mosaic, the problem is considerably well posed. Fig.2(d) show four well distributed pairs of corresponding point features in the mosaic image of two PTZ cameras viewing a planar scene. Fig.2(e) shows the slave-camera view of fig.2(d)(*top*) as seen from the master-camera view of fig.2(d)(*bottom*).

4 Cooperative target tracking

The homographies described in the previous section are now exploited to cooperatively track a target moving in a wide area. The image to world homography $\mathtt{H}$ is used to compute the image sensor likelihood for sequential target tracking in the master camera and the image to image homography $\mathtt{H}'$ is used for camera slaving by computing the homography $\mathtt{T}_k$ of eq.1 (i.e. to transfer imaged target position from the master to the slave camera).

In this section it is shown how to compute the time variant homography $\mathtt{H}_k$. We adopt a SIFT based matching approach to detect the relative location of the current image wrt the reference image: at each time step we extract the SIFT features from the current image and match with those extracted from the reference frame obtaining a set of points' pairs. The SIFT features extracted in the reference image can be considered as visual landmarks. Once visual landmarks are matched to the current view, the registration errors between these points are used to drive a particle filter with state the parameters defining $\mathtt{H}_k$. This allows to stabilize the recovered motion, characterize the uncertainties and reduce the area where matches are searched. Moreover, because the keypoints are detected in scale–space, the scene does not necessarily have to be well–textured which is often the case of planar man–made scene.

Tracking using SIFT Visual Landmarks Let us denote with $\mathtt{H}_k$ the homography between the PTZ camera reference view and the frame grabbed at time step k. What we want to do is to track the parameters that define the homography $\mathtt{H}_k$, using a bayesian recursive filter. Under the assumptions we made, the homography of eq.2 is completely defined once the parameters ψ_k, ϕ_k, and f_k are known, we used this model to estimate the homography $\mathtt{H}_k$ relating the reference image plane $\mathbf{\Pi}'$ with the current image at time k (see fig.1). Thus we adopt the state vector $\mathbf{x}_k$, which defines the camera parameters at time step k: $\mathbf{x}_k = (\psi_k, \phi_k, f_k)$. We use a particle filter to compute estimates of the camera parameters in the state vector. Given a certain observation $\mathbf{z}_k$ of the state vector at time step k, particle filters build an approximated representation of the posterior pdf $p(\mathbf{x}_k|\mathbf{z}_k)$ through a set of weighted samples $\{(\mathbf{x}_k^i, w_k^i)\}_{i=1}^{N_p}$ (called "particles"), where the weights sum to 1. Each particle is thus an hypothesis on the state vector value, with a probability associated to it. The estimated value of the state vector is usually obtained through the weighted sum of all the particles.

As any other bayesian recursive filter, the particle filter algorithm requires a probabilistic model for the state evolution between time steps, from which a prior pdf $p(\mathbf{x}_k|\mathbf{x}_{k-1})$ can be derived, and an observation model, from which a likelihood $p(\mathbf{z}_k|\mathbf{x}_k)$ can be derived. Basically there is no prior knowledge of the control actions that drive the camera through the world, so we adopt a simple random walk model as a state evolution model. This is equivalent to assume the actual value of the state vector to be constant in time and rely on a stochastic noise $\mathbf{v}_{k-1}$ to compensate for unmodeled variations: $\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{v}_{k-1}$. $\mathbf{v}_{k-1} \sim \mathcal{N}(0, \mathbf{Q})$ is a zero mean Gaussian process noise with covariance matrix $\mathbf{Q}$ accounting for camera maneuvers.

The way we achieve observations $\mathbf{z}_k$ of the actual state vector value $\mathbf{x}_k$ is a little more complex and deserves a few more explanations. Let us denote with $\mathcal{S}_0 = \{s_0^j\}_{j=0}^{N}$ the set of SIFT points extracted from the reference view of the PTZ camera (let us assume for the moment a single reference view), and with $\mathcal{S}_k = \{s_k^j\}_{j=0}^{N'}$ the set of SIFT points extracted from the frame grabbed at time step k.

From $\mathcal{S}_0$ and $\mathcal{S}_k$ we can extract pairs of SIFT points that match (through their SIFT descriptors) in the two views of the PTZ camera. After removing outliers from this initial set of matches through a RANSAC algorithm, what remains can be used as an observation for the particle filter. In facts, the set of remaining $\tilde{N}$ pairs: $\mathcal{P}_k = \left\{(s_0^1, s_k^1), ..., (s_0^{\tilde{N}}, s_k^{\tilde{N}})\right\}$ implicitly suggests a homography between the reference view and the frame at time step k, one that maps the points $\{s_0^1, ..., s_0^{\tilde{N}}\}$ into $\{s_k^1, ..., s_k^{\tilde{N}}\}$. Thus, there exist a triple $(\tilde{\psi}_k, \tilde{\phi}_k, \tilde{f}_k)$ which, in the above assumptions, uniquely describes this homography, and that can be used as a measure $\mathbf{z}_k$ of the actual state vector value. To define the likelihood $p(\mathbf{z}_k|\mathbf{x}_k^i)$ of the observation $\mathbf{z}_k$ given the hypothesis $\mathbf{x}_k^i$ we take into account the distance between the homography $\mathtt{H}_k^i$ corresponding to $\mathbf{x}_k^i$ and the one associated to the observation $\mathbf{z}_k$:

$$p(\mathbf{z}_k|\mathbf{x}_k^i) \propto e^{-\frac{1}{\lambda}\sqrt{\sum_{j=1}^{M}\left(\mathtt{H}_k^i \cdot s_0^j - s_k^j\right)^2}} \tag{5}$$

where $\mathtt{H}_k^i \cdot s_0^j$ is the projection of s_0^j in the image plane of frame k through the homography $\mathtt{H}_k^i$, and λ is a normalization constant.

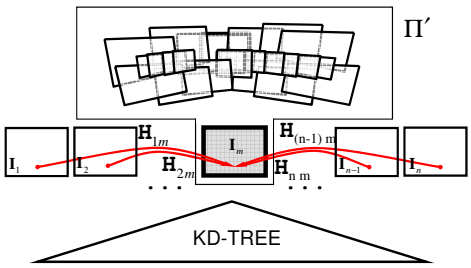


Fig. 3. Each landmark in the database has a set of descriptors that corresponds to location features seen from different vantage points. Once the current view of the PTZ camera matches an image $\mathbf{I}_l$ in the database, the inter-image homography $\mathtt{H}_{lm}$ is used to transfer the current view into the reference plane $\mathbf{\Pi}'$.

It is worth to note that the SIFT points on the frame k do not need to be computed upon the whole frame. In facts, after the particle filter prediction step it is possible to reduce the area of the image plane where the SIFT points are computed to the area where the particles are propagated. This reduces the computational load of the SIFT points computation and of the subsequent matching with the SIFT points of the reference image.

To increment robustness of the recursive tracking described above, during a learning stage a database of the scene feature points is build. SIFT keypoints extracted to compute the mosaic are merged into a large KD-Tree together with the estimated mosaic geometry. The match for a SIFT feature extracted from the current frame is searched according to the Euclidean distance of the descriptor vectors. The search is performed so that bins are explored in the order of their closest distance from the query description vector, and stopped after a given number of data points has been considered [11].

Once the image $\mathbf{I}_l$ closest to the current view $\mathbf{I}_k$ is found the homography $\mathtt{G}$ relating $\mathbf{I}_k$ to $\mathbf{I}_l$ is computed at run time with RANSAC. The homography $\mathtt{H}_{lm}$ that relates the image $\mathbf{I}_l$ with the mosaic plane $\mathbf{\Pi}'$ retrieved in the database is used to finally compute the likelihood. Eq.5 becomes:

$$p(\mathbf{z}_k|\mathbf{x}_k^i) \propto e^{-\frac{1}{\lambda}\sqrt{\sum_{j=1}^{M}\left(\mathtt{H}_k^i \cdot s_0^j - \mathtt{H}_{lm}\cdot\mathtt{G}\cdot s_k^j\right)^2}} \tag{6}$$

As shown in fig.3 the image points of the nearest neighbor image $\mathbf{I}_l$ wrt to current view $\mathbf{I}_l$ and the current view (i.e. the query to the database) are projected in $\mathbf{\Pi}'$ to compute the likelihood of eq.6. In particular $\mathbf{I}_m$ in the figure is the reference image used to compute the mosaic.

5 Experimental Results

In order to test the validity of the presented method we have acquired 40 images with two IP PTZ-camera Sony SNC-RZ30 in a master-slave configuration. Fig.2(a) shows the images used taken from one PTZ camera. The input images are taken at a resolution of 736×544 pixels. The images from the other camera are not shown. The images has been captured at the minimal zoom of the device and with a pan and tilt angle increment of respectively $\psi = 27.14$ and $\phi = 10.0$, so as to have some overlap between images. For a given image, matches are searched only at the 8 neighbor images in the grid (apart for those images in the border of the grid). Once correspondences are formed through RANSAC, the homographies parameterized by eq.2 are computed and are successively refined by non linear minimization through bundle adjustment. The RANSAC strategy successful rejects outliers, even if several moving objects are present in the scene.

The image in the second row, fifth column in the grid of fig.2(a) (i.e. the reference image) is used to stitch the mosaic so as to avoid degeneracies in the estimation of the image of the absolute conic ω. Parallel lines have been manually extracted by following the imaged linear boundaries. Fig.2(b) shows the features point over the image boundaries in the image mosaic representing the pairs of mutually orthogonal parallel lines used to compute the vanishing points. The lines are fitted by orthogonal regression. In the same figure it is also shown the orthocenter $\mathbf{p}$ of the vanishing point triangle and the reference image is indicated with a rectangle. Fig.2(c) shows the image mosaic transformed by the rectifying homography of eq.3. The global Euclidean structure of the 3D world plane is recovered.

In extended areas, obtaining a ground truth homography is operatively difficult to be made with a rule. This prevents an extensive statistical evaluation in real scenario. For this reason we preferred to compare the method described with a conventional method in a real scene. We measured four world point coordinates which project onto the input image located in the first row, seventh column of the grid in fig.2(a). Hence we compute the world to image homography which is used to rectify the mosaic. The world points are distant no more than 15 meters each other. Their positions are computed by measuring the inter-distance between the 3D marker and then solving for the 3D coordinates. The result is

shown in fig.2(f), it is evident that the lines of the street in the courtyard are no more parallel after the rectification.

The right angle of the sidewalk of the courtyard can be used to validate the accuracy. The angle is measured to be nearly 90° by the rectification using eq.3, while the angle in fig.2(f) measures 76°. This can be explained by the fact that the world to image homography is quite accurate locally where the measurements are taken, while as we move from these position errors increase. To further appreciate this behavior fig.2(g) shows a regularly grid of equispaced points backprojected over the imaged planar region in the mosaic. In particular it is also shown the reference image (the rectangle) and the imaged world points used to compute the homography with the conventional method. Fig.2(h) shows a global view of same figure. The imaged grid is very inaccurate outside the reference image. The homography computed by the measured local features does not give good global results also because large outdoor scene areas may deviate from being planar.

For testing purposes, a simple algorithm has been developed to automatically track a single target using the recovered homographies. The target is localized with the wide angle stationary camera using background subtraction and its motion within the image is modeled using an Extended Kalman Filter (EKF). The observation model is obtained by the linearization of eq.4. Images are used to compute respectively the image to world homography for the master and the inter-image homography relating the mosaic plane of the two PTZ cameras. Because of the limited extension of the monitored area, a wide angle view of the master camera suffice to track the target. The feature points of the slave camera images are used to build the database of SIFT for camera tracking. In fig.4(a) are shown some frames extracted from an execution of the proposed system: on the top row is shown the position of the target observed with the master camera, on the bottom the frames of the slave camera view. The particles show the uncertainty on the position of the target. Since the slave camera does not explicitly detects the target, the background color similar to the foreground color does not influence the estimated localization of the target.

A quantitative result for the estimated camera parameters is depicted in fig.4(b). It can be seen that increasing the focal length usually causes a significant increase in the variance also, which means that the estimated homography between the two cameras become more and more inaccurate. Observing in detail the particle filter for camera tracking, we noticed from our experiments that the uncertainty increase with the zoom factor. This is caused by the fact that features at high resolution that match with those extracted from the reference image obviously decrease when zooming, causing SIFT match to be less accurate. However this error remains bounded below certain zoom factors, about 70%.

6 Summary and Conclusions

In this paper we have shown how to combine single view geometry and planar mosaic geometry in order to define and solve the two basic building blocks defining PTZ camera networks. Those are the world to image and the inter-image

homography. The main virtue of our results lies in the simplicity of the method. Future research will investigate the joint optimization of the single view geometry together with the mosaic registration in terms of the IAC parameterization. However the most interesting direction (currently under investigation) will use the presented framework to compute a selective attention strategy, aimed to track multiple targets by tasking the sensors in the network.

References

1. A. Bartoli, N. Dalal, and R. Horaud. Motion panoramas. *Computer Animation and Virtual Worlds*, 15:501–517, 2004.
2. Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, and Hasegawa. A system for video surveillance and monitoring: Vsam final report. *Technical report CMU-RI-TR-00-12, Robotics Institute, Carnegie Mellon University*, May 2000.
3. C. Colombo, A. D. Bimbo, and F. Pernici. Metric 3d reconstruction and texture acquisition of surfaces of revolution from a single uncalibrated view. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 27(1):99–114, 2005.
4. C. J. Costello, C. P. Diehl, A. Banerjee, and H. Fisher. Scheduling an active camera to observe people. *Proceedings of the 2nd ACM International Workshop on Video Surveillance and Sensor Networks*, pages 39–45, 2004.
5. J. Davis and X. Chen. Calibrating pan-tilt cameras in wide-area surveillance networks. *In Proc. of ICCV 2003*, 1:144–150, 2003.
6. L. de Agapito, E. Hayman, and I. D. Reid. Self-calibration of rotating and zooming cameras. *International Journal of Computer Vision*, 45(2), November 2001.
7. A. del Bimbo and F. Pernici. Distant targets identification as an on-line dynamic vehicle routing problem using an active-zooming camera. *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05) in conjunction with ICCV, Beijing, China*, pages 15–21, October 2005.
8. R. Hartley. Self-calibration from multiple views with a rotating camera. *in Proc. European Conf. Computer Vision*, pages 471–478, 1994.
9. R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.
10. D. Liebowitz, A. Criminisi, and A. Zisserman. Creating architectural models from images. In *Proc. EuroGraphics*, volume 18, pages 39–50, September 1999.
11. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.
12. S. Sinha and M. Pollefeys. Towards calibrating a pan-tilt-zoom cameras network. *P. Sturm, T. Svoboda, and S. Teller, editors, OMNIVIS*, 2004.
13. P. P. Sturm and Y. Wu. Euclidean structure from n≥2 parallel circles: Theory and algorithms. *In Proc. of the 9th European Conference on Computer Vision (ECCV'2006).*, pages 238–252, 2006.
14. T. Svoboda, H. Hug, and L. V. Gool. Viroom – low cost synchronized multicamera system and its self-calibration. *In Pattern Recognition, 24th DAGM Symposium, number 2449 in LNCS,*, pages 515–522, September 2002.
15. X. Zhou, R. Collins, T. Kanade, and P. Metes. A master-slave system to acquire biometric imagery of humans at a distance. *ACM SIGMM 2003 Workshop on Video Surveillance*, pages 113–120, 2003.

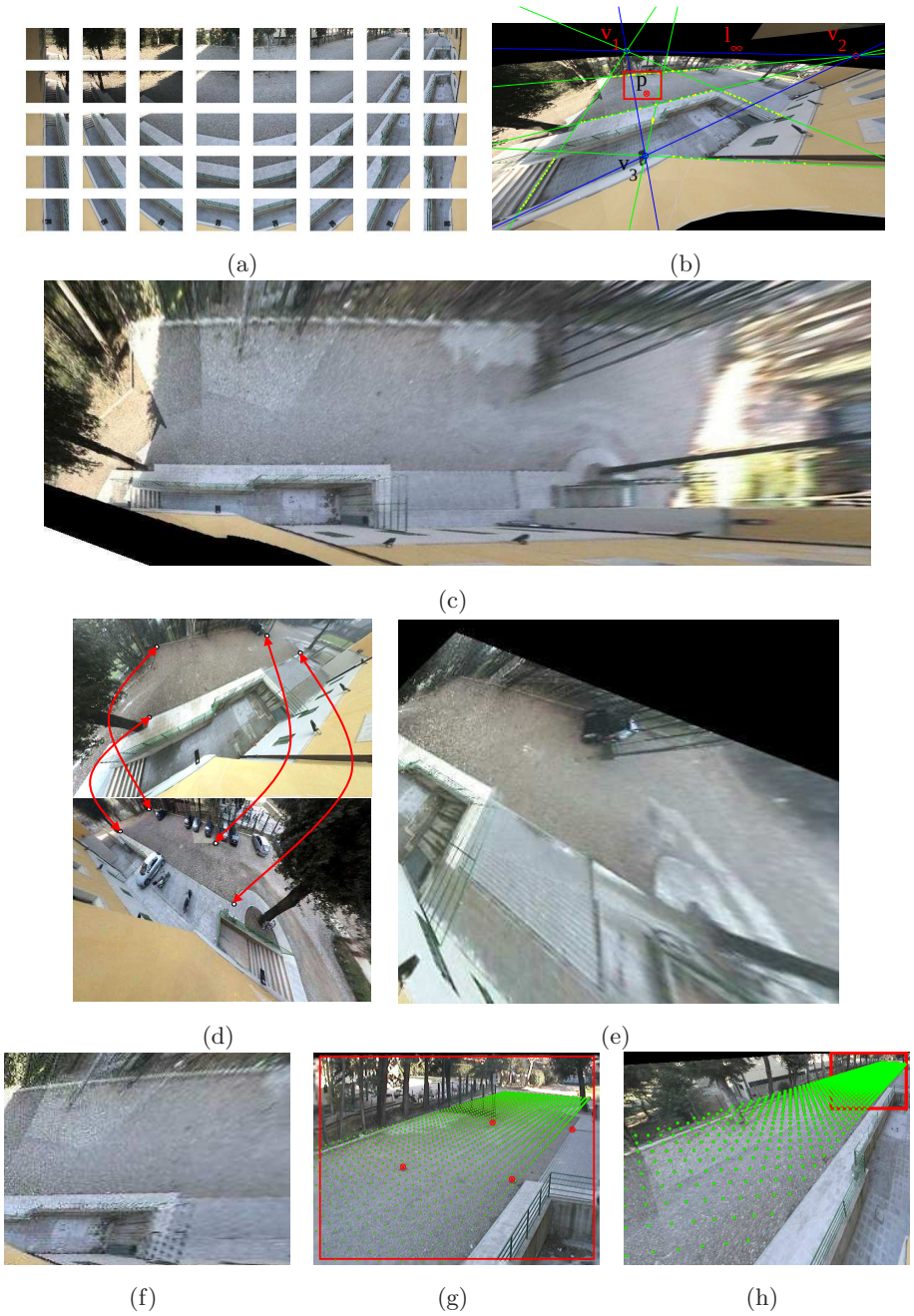
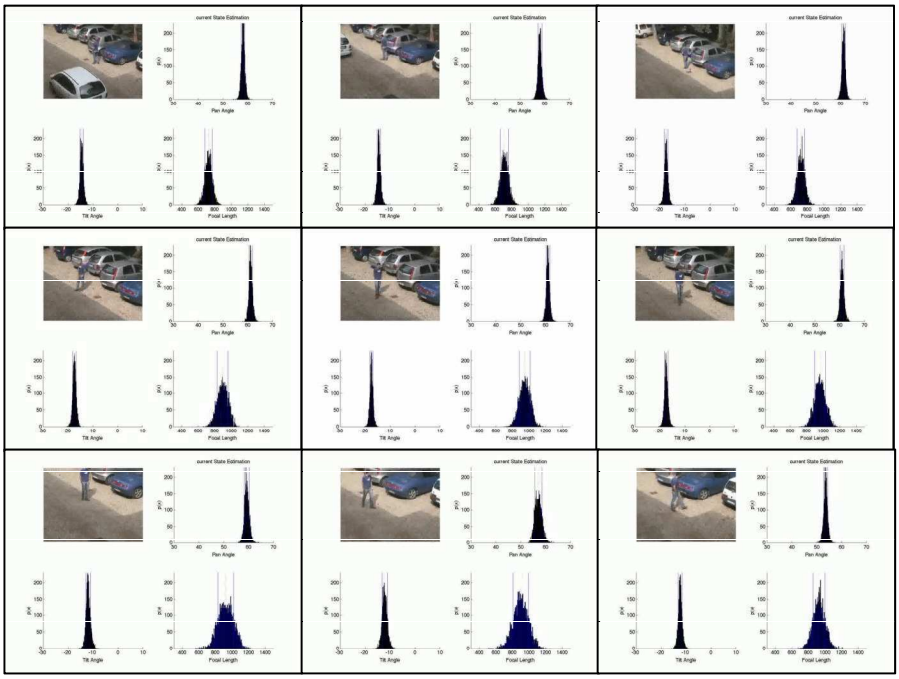


Fig. 2. (a): The grid of 40 input images captured at the minimal zoom of the device and with a pan and tilt angle increment of respectively $\psi = 27.14$ and $\phi = 10.0$. (b): The planar mosaic with superimposed the vanishing point triangle $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$. (c): The rectified mosaic. In this picture is shown the rectification through the homography of eq.3. The global Euclidean structure of the 3D world plane is recovered. (d): Four well spaced pairs of corresponding points used compute the inter-image homographies relating the two PTZ camera mosaics. (e): The slave-camera field of regard view as seen from the master-camera field of regard. (f): Mosaic planar rectification using 3D known measures. The image used to compute the world to image homography is also used as a reference image to stitch the mosaic. (g): The reference image (rectangle) in the mosaic. The figure also shows the backprojection of the four known 3D points, and the backprojection of a grid of points. (h): A global view of fig.(g) superimposed in the mosaic (top-right rectangle). The grid is inaccurate outside the reference image.

(a)



(b)

Fig. 4. (a): On the top row, the master slave camera tracking a human target. The world to image homography is estimated from the vanishing points in the mosaic image and used as observation model in an Extended Kalman Filter. On the bottom row, the slave camera viewing the target tracked from the master. The particles show the joint position uncertainty of the target and the slave camera. (b): Nine frames showing the probability distributions (histograms) of the slave camera parameters. In particular in each frame are shown (left-to-right top-to-bottom) the current view, the pan, the tilt and the focal length distributions. As one would expect, uncertainty increase with zoom factor.