Computer Vision and Image Understanding 114 (2010) 611–623

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Exploiting distinctive visual landmark maps in pan–tilt–zoom camera networks

A. Del Bimbo, F. Dini, G. Lisanti, F. Pernici *

Media Integration and Communication Center (MICC), University of Florence, Viale Morgagni 65, Florence 50134, Italy

ARTICLE INFO

Article history:
Received 11 January 2009
Accepted 5 January 2010
Available online 21 January 2010

Keywords:
Pan–tilt–zoom camera
Rotating and zooming camera
Camera networks
Distinctive keypoints
Tracking
Master–slave configuration

ABSTRACT

Pan–tilt–zoom (PTZ) 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. One method to achieve such behavior is to 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 can thus point towards the targets in high resolution.

In this paper we describe a novel framework exploiting a PTZ camera network to achieve high accuracy in the task of relating the feet position of a person in the image of the master camera, to his head position in the image of the slave camera. Each camera in the network can act as a master or slave camera, thus allowing the coverage of wide and geometrically complex areas with a relatively small number of sensors.

The proposed framework does not require any 3D known location to be specified, and allows to take into account both zooming and target uncertainties. Quantitative results show good performance in target head localization, independently from the zooming factor in the slave camera. An example of cooperative tracking approach exploiting with the proposed framework is also presented.

1. Introduction

In realistic surveillance scenarios, it is impossible for a single camera sensor either fixed or with pan–tilt–zoom (PTZ) capabilities to monitor outdoor wide areas entirely so as to be able to detect and track moving entities and discover interesting events. In fact, small changes of the viewpoint can determine large differences in the appearance of the moving entities due to illumination, cast shadows and (self-)occlusions and therefore drastically impact the performance of object detection and tracking as well as of recognition. To solve this problem, camera networks are employed to acquire multiple views of the entities from different viewing angles and therefore recover the information that might be missing if observed from a single viewing direction. Fixed cameras are generally adopted, being sufficiently simple to compute their relative spatial relationships [24]. Although fixed camera networks has been successfully applied in real application contexts, nevertheless they still suffer from the inherent problem of sensor quantization. In fact, fixed optics and fixed sensor resolution can make the structure of far-away entities similar to the texture of near-field entities. Super-resolution algorithms [30] applied to low resolution video frames do little to improve video quality.

Instead, effective solution to this problem can be obtained from the combination of a fixed camera with a PTZ camera working in a cooperative way. The two cameras are typically settled in a master–slave configuration [49]: the master camera is kept stationary and set to have a global view of the scene so as to permit to track several entities simultaneously. The slave camera is used to follow the target trajectory and generate close-up imagery of the entities driven by the transformed trajectory coordinates, moving from target to target and zooming in and out as necessary. In the most general case, this master–slave configuration can be exploited in a PTZ camera network, where several slave PTZ cameras can be controlled from one or several master PTZ camera(s) to follow the trajectory of some entities and generate multi-view close-up imagery in high resolution. In this framework, each master camera operates as if it was a reconfigurable fixed camera. An important capability of PTZ camera networks, particularly useful in biometric recognition in wide areas, is that of focusing on interesting human body parts such as head [38].

However, the working implementation of PTZ camera networks poses much more complex problems to solve than classical stationary camera networks. Assuming that all the cameras observe a planar scene, the image relationships between the camera image planes undergo a planar time-variant homography. But background appearance is not stationary and camera parameters change through time as the PTZ cameras pan, tilt and zoom, so making it difficult to compute their relative spatial positions. Estimating the time-variant image-to-image homography between a fixed master and a slave PTZ camera in real-time is also challenging. Occlusions, sensor quantization and foreshortening effects significantly limit the area of the PTZ camera view where to search

* Corresponding author.
E-mail addresses: delbimbo@dsi.unifi.it (A.D. Bimbo), dini@dsi.unifi.it (F. Dini), lisanti@dsi.unifi.it (G. Lisanti), pernici@dsi.unifi.it (F. Pernici).

1077-3142/$ - see front matter © 2010 Elsevier Inc. All rights reserved.
doi:10.1016/j.cviu.2010.01.007

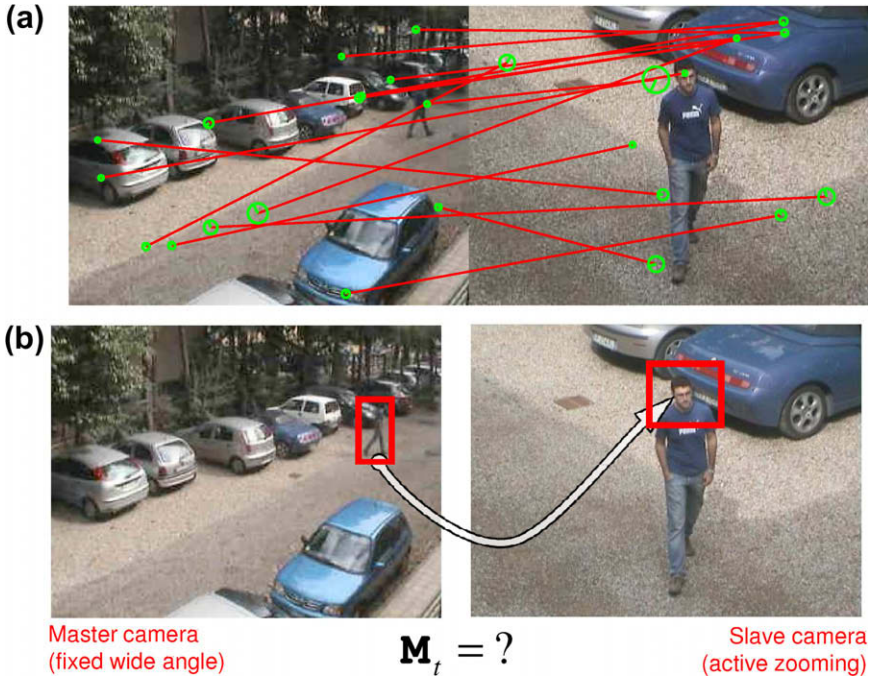

Fig. 1. Principal problems to solve with PTZ camera networks: (a) failure of SIFT matching; image sensor quantization and occlusions significantly impair the search for correspondences. (b) Estimation of the time-variant transformation $\mathbb{M}_t$ that maps feet to head from the master to the slave view. Left: Wide angle master camera view in which a person is detected. Right: Close up view from the slave camera.

for feature matches with the view of the master camera, making therefore difficult to compute the corresponding homography. In addition, the magnification factor achieved by zooming cameras can determine a large variation of the image structure, so limiting the matching performance. Fig. 1 exemplifies the principal problems to be solved in this framework.

In this paper, we discuss a novel solution for the effective implementation and real-time operation of PTZ camera networks. The approach that is proposed exploits a prebuilt map of visual 2D landmarks of the wide area to support multi-view image matching. The landmarks are extracted from a finite number of images taken from a non calibrated PTZ camera, in order to cover the entire field of regard.[1] Each image in the map also keeps the camera parameters at which the image has been taken. At run-time, features that are detected in the current PTZ camera view are matched to those of the base set in the map. The matches are used to localize the camera with respect to the scene and hence estimate the position of the target body parts. Fig. 2 shows the main components of our system. A discussion of the motivations and basic ideas underlying the approach followed has been presented in some detail in [16].

We provide several new contributions in this research:

- A novel uncalibrated method to compute the time-variant homography, exploiting the multi-view geometry of PTZ camera networks. Our approach avoids drifting and does not require calibration marks [23] or manually established pan–tilt correspondences [49].
- The target body part (namely the head) is localized from the background appearance motion of the slave zooming camera. Head or face detection and segmentation are not required. Differently from [38] our solution explicitly takes into account camera calibration parameters and their uncertainty.
- Differently from [1,49], where a PTZ camera and a fixed camera are set with a short baseline so as to ease feature matching between the two fields of view, in our solution we define a general framework for arbitrary camera network topology. In this framework, any node of the network sharing a common field of view can exploit the master–slave relationship between cameras.

In the following we first provide an overview of the related work in Section 2. Hence, in Section 3, PTZ camera networks with master–slave configuration are defined in terms of their relative geometry and functionality. The details of map building process are presented in Section 4. Camera pose tracking and sensor slaving are presented in Section 5. System performance is discussed in Section 6, followed by final remarks.

2. Related work

Sensor slaving is a relatively simple practice provided that both the master and the slave camera are calibrated with respect to a local 3D terrain model [9]. Camera calibration allows to transfer 3D object locations onto the camera image plane and therefore use this information to steer the pan tilt and zoom of the slave sensor in the appropriate direction.

Several methods have been published in the literature to perform calibration of PTZ cameras. Early works have concentrated the attention on internal camera parameters estimation, with no support for on-line dynamic calibration. A first significant work for active zoom lens calibration was published by Willson et. al. [47]. They considered indoor scenes and controlled environments and used calibration targets. In this framework, they exploited the fact that active zooming cameras, if stationary, play the same role of fixed cameras and therefore standard methods for fixed cameras still apply for their calibration. However their method lacks the needed flexibility to be used in outdoor wide areas with moving PTZ cameras.

[1] The camera field of regard is defined as the union of all fields of view over the entire range of pan and tilt rotation angles and zoom values.

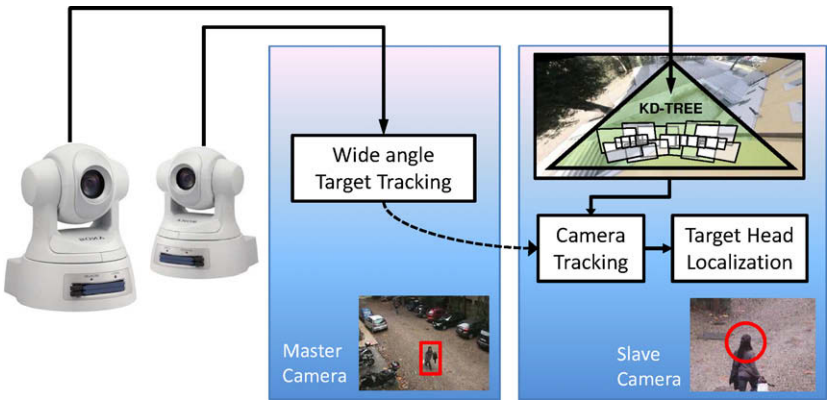

Fig. 2. Left: Off the shelf PTZ cameras. Right: Components of our system and their connections, executed on-line for each frame of the video stream.

Other researchers have proposed methods for self-calibration with no use of calibration marks. The method proposed in [19] permits self-calibration of a PTZ camera calculating the homographies induced by the rotation and zooming of the PTZ camera. In [15], the authors followed the same approach and analyzed the effects of imposing simplifying 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 varying in reality. In [39], a complete evaluation of the method of [19] was performed using a large set of images. The 3D scene under observation was reconstructed from the internal calibrations of two PTZ cameras using the mosaic images as a stereo pair.

Objects moving around in the observed scene have also been used for PTZ camera self-calibration. In [13,44] LEDs have been employed to this end. 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. This approach is nevertheless too cumbersome and needs too accurate synchronization to be used in generic outdoor environments.

Of more general application is instead the use of walking people to perform weak camera calibration [40]. In [6,27,32] camera calibration is obtained exploiting single view geometry and the vanishing point derived from image features of the moving objects. Although the method is appropriate for PTZ camera (self-)calibration in outdoor environments the parallel lines used to compute the vanishing points must be viewed under strong perspective, that is a condition that does not apply at moderate zooming. Moreover the measured features are computed by blob segmentation and are too noisy to permit reliable estimation of the geometric models.

After the VSAM project [9], new methods have been proposed for calibrating PTZ cameras with simpler and more flexible approaches with less intensive off-line processing. These methods are more suitable for outdoor environments and explicitly address high resolution zooming of targets at a distance. The master–slave configuration includes two cameras [49,1,38,11,3,34,17] or several cameras [29,42,18,10].

Among them, the solutions proposed in [49,1] do not require direct calibration but impose some restrictions in the setup of the cameras. The viewpoints between the master and slave camera are assumed to be nearly identical so as to ease feature matching. In [49], a linear mapping is used that is computed from a look-up table of manually established pan and tilt correspondences. In [1], a look-up table is employed that also takes into account camera zooming. In [38], it is proposed a method to link the foot position of a moving person in the master camera sequence with the same position in the slave camera view. The methods proposed by [21,23,18,42] require instead direct camera calibration, with a moving person and calibration marks.

Real-time estimation of camera's position and orientation relative to some geometric representation of its surroundings using visual landmarks has been proposed by other authors, following the Monocular Simultaneous Localization and Mapping (monoSLAM) approach. Similar in principle to Structure from Motion Techniques (SfM), the SLAM approach performs on-line recursive recovery of the scene, while exploiting the correlations between the observations of the camera and the scene entities. Typically internal camera parameters are known in advance [14], while in SfM they are estimated jointly with 3D structure [20]. Scale-invariant feature transform (SIFT) and matching based on best-bin first k-d tree search [31] were used in [37] for robot localization and mapping to find the visual landmarks and establish their correspondences. However, due to the complexity of the SIFT descriptor the number of feature points that can be handled simultaneously is low, thus limiting the reliability of the method. Improvements in the scalability and robustness of data association with SIFT were suggested in [7].

Recent research achievements of effective local image descriptors [48,43,33,22,4,36] have nevertheless offered the opportunity of new improvements for this approach and its operation in full real-time: commercial SIFT implementations run even at frame rate using careful coding and processor extensions; SURF descriptors [4] achieve one order of magnitude of performance improvement, by exploiting integral images; FAST corner detection [36] achieves frame rate operation, although it is less robust to motion blur. Real-time global optimization with on-line SLAM applied over a number of frames has been implemented in several systems [8,25,41] considering cameras with pre-calculated internal calibration. However all these systems only consider the case of camera panning and tilting with smooth motion. No zooming camera operation is considered so that their applicability is restricted to simple and special cases.

3. Geometric relationships and camera model

3.1. Basic geometric relationships

If all the cameras in a network have an overlapping field of view (i.e. they are in a fully connected topology), they can be set in a master–slave relationship pairwise. According to this, given a network of M PTZ cameras $\mathbf{C}_i$ viewing a planar scene, $\mathcal{N} = \{\mathbf{C}_i^s\}_{i=1}^M$, at any given time instant each camera can be in one of two states $s_i \in \{\text{master}, \text{slave}\}$. The network can be therefore in one of $2^M - 2$ possible state configurations. All cameras in MASTER, or all

cameras in SLAVE state cannot be defined. Several cameras can instead be either in the MASTER or in the SLAVE state. If several master cameras are defined, they permit multiple observations of the same target from different viewpoints and therefore more accurate measurements and increased tracking accuracy. In this case, one slave camera can be sufficient to observe the area with accurate foveation. One single master camera and several slave cameras permit instead to capture high resolution images of moving targets from several viewpoints.

Fig. 3 shows the basic geometry of a PTZ camera network with the pairwise relationships to perform sensor slaving. The homographies H_{12}, H_{13}, H_{23} put in pairwise relationship the reference planes Π_1, Π_2, Π_3 of camera $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$. The homographies H_t^1, H_t^2 and H_t^3 relate instead the reference image planes Π_1, Π_2, Π_3 with the current image plane at time t. If the target $\mathbf{X}$ is tracked by $\mathbf{C}_1$ (acting as MASTER) and followed in high resolution by $\mathbf{C}_2$ (acting as zooming SLAVE), the imaged coordinates of the target are first transferred from Π_1 to Π_2 through H_{12} and hence from Π_2 to the current zoomed view of $\mathbf{C}_2$ through H_t^2. Referring to the general case of M distinct cameras, once H_t^k and H_{kl}, $k \in 1,\ldots,M$, $l \in 1,\ldots,M$ with $l \neq k$ are known, the imaged location of a moving target tracked by a master camera $\mathbf{C}_k$ can be transferred to the zoomed view of a slave camera $\mathbf{C}_l$ according to:

$$\mathrm{T}_t^{kl} = \mathrm{H}_t^l \cdot \mathrm{H}_{kl} \tag{1}$$

3.2. PTZ camera geometry model

When cameras are used outdoor or in very large environments, the deviation of the camera optical center (the nodal point) is negligible compared to the average distance of the observed features [35]. According to this, we consider the pin-hole camera model projecting the three-dimensional world onto a two-dimensional image, with fixed principal point without modeling the radial distortion; we also assume that the camera rotates around its optical center with no translation and the pan and tilt axes intersect each other.

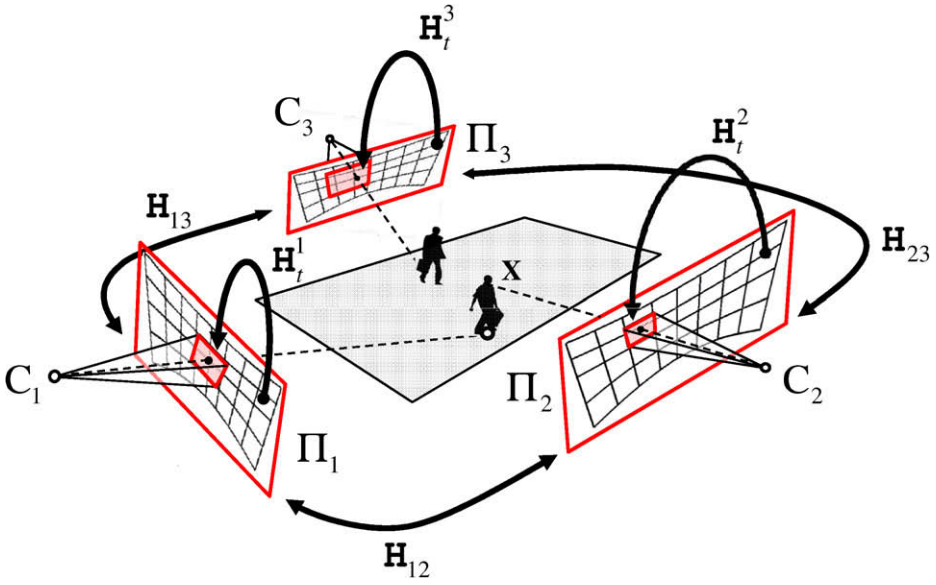

Fig. 3. The pairwise relationships between PTZ cameras in master–slave configuration for a sample network with three PTZ cameras.

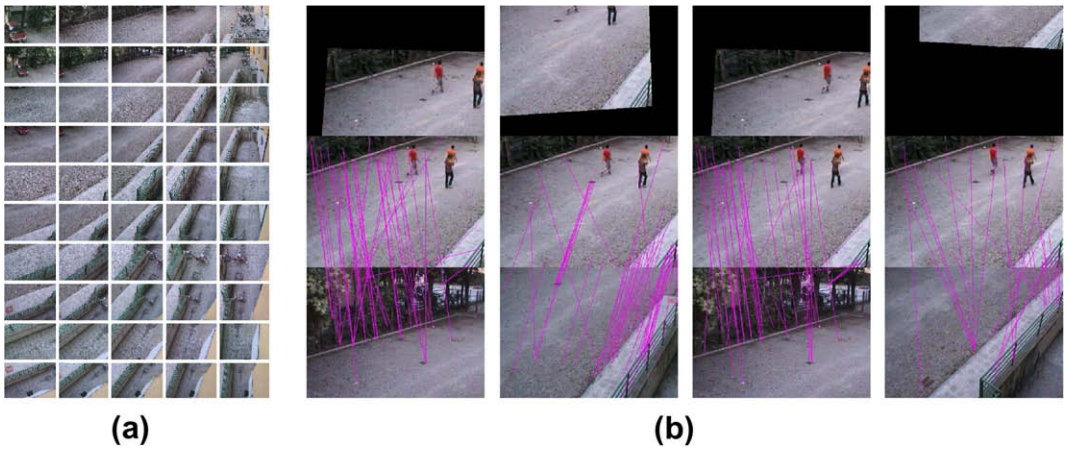

Fig. 4. SURF keypoints detection and matching between PTZ camera views and pre-stored scene map built from reference images. (a) The grid of reference images used to build the map. (b) Results of on-line SURF keypoints detection and matching (consecutive frames taken at 15 frame per second shown): current frames from the PTZ camera image (central row); nearest neighbor image found in the map for each PTZ frame (bottom row); current frame warped to the nearest neighbor image found in the map according to the homography estimated (top row). Lines indicate the matches of SURF keypoints used to estimate the homographic warping.

For a generic image i generated by a camera C, projection can be therefore modelled as $\mathtt{P}_i = [\mathtt{K}_i\mathtt{R}_i\ 0]$, where $\mathtt{K}_i$ is the 3×3 matrix that contains the intrinsic parameters of the camera, and $\mathtt{R}_i$ is the 3×3 matrix that defines the camera orientation; the equal sign denotes equality up to a scale factor. As in [20], it is possible to derive the inter-image homography, between image i and image j generated by the same camera, as: $\mathtt{H}_{ji} = \mathtt{K}_j\mathtt{R}_{ji}\mathtt{K}_i^{-1}$.

For PTZ cameras, due to their mechanics, it is possible to assume that there is no rotation around the optical axis, i.e. $\theta = 0$. We will also assume with good approximation that the principal point lies at the image center, the pan–tilt angles between spatially overlapping images are small and the focal length does not change 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, [2]. 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 $\mathtt{H}_{ji}$.

Using the image-to-image homography of Eq. (2) makes runtime matching and minimization in PTZ camera networks much simpler than with the full 8 DOF homography. Even if the calibration parameters are not accurate, it is nevertheless possible to create a wide single view of the entire scene (i.e. a planar mosaic) from a finite number of reference images taken with one PTZ camera at different pan, tilt and zoom settings so as to cover the entire field of regard, still 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 and used to localize the PTZ camera views at run-time.

4. Building a global map of the scene

In our approach, images of the scene taken from a non calibrated PTZ camera at different values of pan, tilt and zoom are collected off-line to build up a global map of the scene under observation, keeping memory of the geometric information at which each image was taken. The construction of the global map of the scene is decoupled from frame-to-frame tracking of the camera pose and focal length so that data association errors are not integrated into the map and more precise tracking is obtained. Scale and rotation invariant keypoints are extracted from each image and used at run-time to match the keypoints extracted from the current frame of the PTZ camera. Finding the right correspondence between the current view and the scene permits to evaluate the homographies in Eq. (1) and localize the camera with respect to the scene. The approach is similar to [26,28], with a few key differences. On one hand we have two parameters for camera pose (i.e. pan and tilt angles) instead of six; on the other hand we have varying internal camera parameters (i.e. focal length).

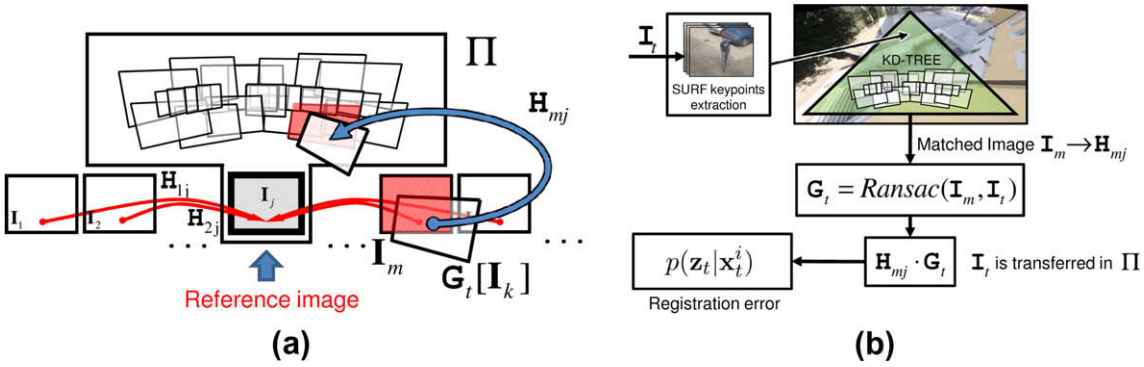

Fig. 5. Once the current view $\mathbf{I}_t$ of the PTZ camera matches an image $\mathbf{I}_m$ in the map through the homography $\mathtt{G}_t$, the inter-image homography $\mathtt{H}_{mj}$ is used to transfer the current view $\mathbf{I}_t$ into the camera reference plane Π. Image $\mathbf{I}_j$ is used as a reference image to build the mosaic map; (a) scene map and its components (off-line) and (b) main processing steps for tracking PTZ camera parameters (on-line).

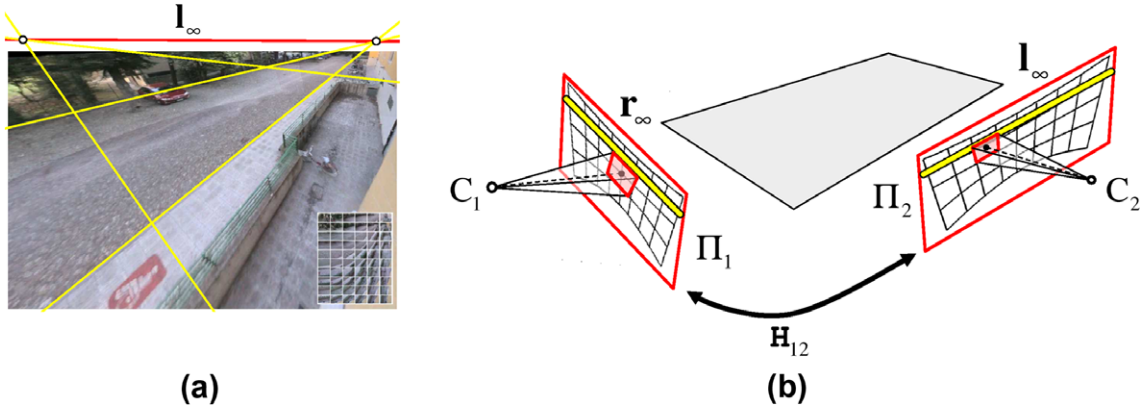

Fig. 6. (a) An example of vanishing line $\mathbf{l}_\infty$ in the camera reference plane computed from two pairs of imaged 3D parallel lines, here shown superimposed in a low resolution mosaic. (b) The vanishing line is computed once in one camera ($\mathbf{r}_\infty$ in $\mathbf{C}_1$), and transferred to the other camera ($\mathbf{l}_\infty$ in $\mathbf{C}_2$) through $\mathtt{H}_{12}$.

Fig. 4 shows an example of this approach. Even if few distinctive keypoints are present, the images in the map taken at different resolution and pan/tilt angles still allow to find the necessary correspondences to localize the camera. SURF keypoints are used as visual landmarks of the scene. The use of SURF features to match the current view of the camera is motivated by the fact that, in the case of outdoor PTZ camera operation, they permit a more precise definition of scale and shape than corners, despite of the fact that blob-like structures are less accurately localized than corners in the image plane. SURF keypoints are in fact invariant to the similarity transformation that small image patches undergo when the PTZ camera rotates about its optical centre. The boundaries of a blob, even if irregular, provide a good estimate of the size (and therefore of the scale) of the blob, so significantly improving the repeatability of detection under scale change. In addition, SURF keypoints do not require that the scene is well-textured, and are robust to motion blur (keypoints are detected in scale–space). However, since image blobs are not accurately localized they cannot be used to estimate the focal length from the homography between two overlapping frames, especially when the focal length increases. According to these considerations, inter-image homographies of Eq. (2) are estimated from the extracted keypoints using bundle adjustment optimization on the reference images as in [45,15]. All the estimated homographies are related back to a reference image (i.e. to the reference plane). In order to support tracking at frame rate of the camera pose and focal length, all the keypoints of the global map and their camera geometry information are stored in a k-d tree. Each keypoint in the k-d tree is associated with the image it comes from, and in its turn, each image is associated with the homography that relates it back to the reference plane, so that quick matching and camera geometry retrieval can be performed on-line with the keypoints observed in the current PTZ camera view.

5. Online master–slave relationship estimation

Matching of current frame keypoints with those in the global map is made according to nearest neighbour search in the feature descriptor space. We followed the Lowe's technique [31] that assumes the 1-NN in some image is a potential correct match, while the 2-NN in the same image is an incorrect match. The final image $\mathbf{I}_m$ is the one that has the highest number of feature matches with respect to the current image $\mathbf{I}_t$. Once the image $\mathbf{I}_m$ is found, the correct homography $\mathtt{G}_t$ relating $\mathbf{I}_t$ to $\mathbf{I}_m$ is computed at run-time using RANSAC (see Fig. 4), exploiting only the features of $\mathbf{I}_m$.

The homography $\mathtt{H}_{mj}$ that relates the image $\mathbf{I}_m$ with the image $\mathbf{I}_j$ in the reference plane Π that is retrieved in the k-d tree hence is used to compute the likelihood to estimate $\mathtt{H}_t$ in Eq. (1). Details are given in Section 5.1. In Section 5.2 we further exploit the refer-

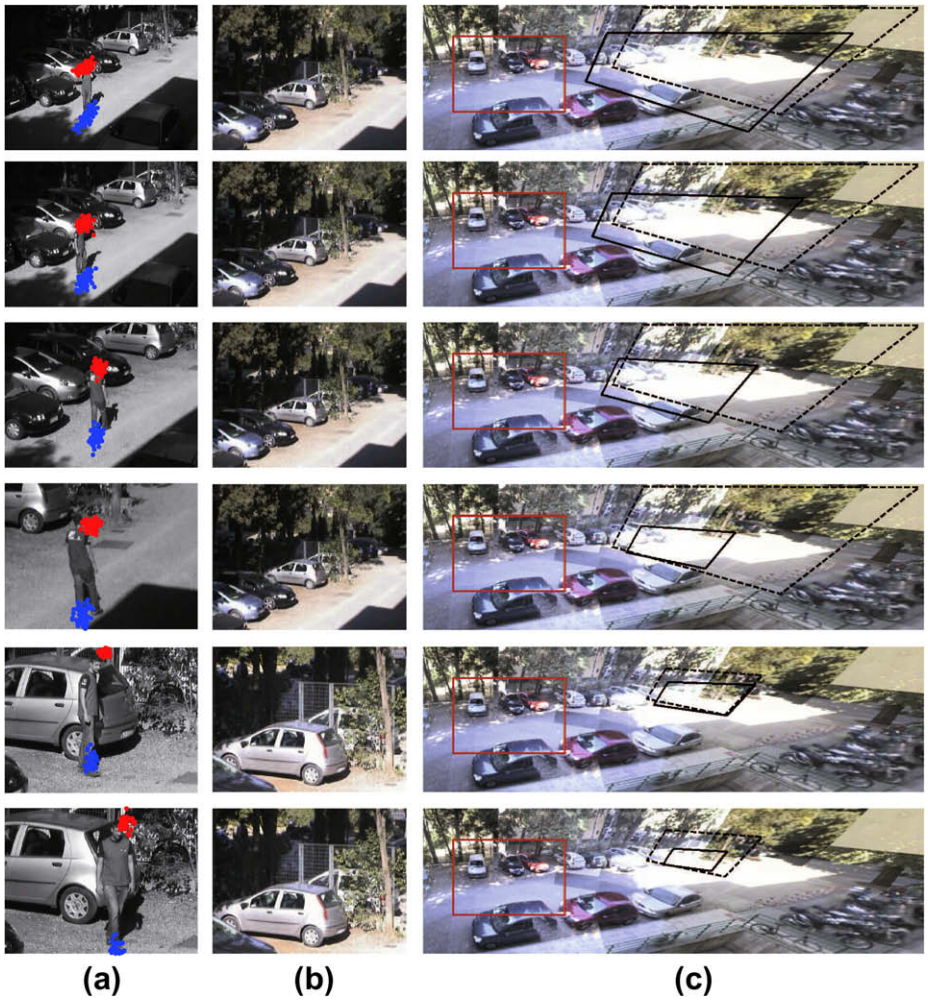

Fig. 7. Six intermediate frames generated by our method for the first sequence: (a) sequence frames (from top to bottom) as taken from the slave camera with superimposed particles. (b) Associated nearest neighbor images retrieved from the map. (c) The slave camera reference plane with superimposed mosaic. The rectangle shows the reference image $\mathbf{I}_j$. The solid and the dashed polygons show respectively the transformed boundary of the current image $\mathbf{I}_t$ with $\widehat{\mathtt{H}}_t$ (i.e. the averaged filter state homography) and the transformed boundary of image $\mathbf{I}_m$ with $\mathtt{H}_{mj}$. Fifth row: A better featured image is automatically selected as the camera zooms in.

ence plane image-to-image homography and the vanishing line to transfer the target position from the master camera to the slave camera and to locate the target's head.

5.1. Slave camera tracking using SURF visual landmarks

In order to track the movements of a slave PTZ camera, we have to track the parameters that define the homography $\mathtt{H}_t$ between the reference plane and the frame grabbed at time t. Under the assumptions made, this homography is completely defined once the parameters ψ_t, ϕ_t, and f_t are known. To this end we use particle filtering to perform the estimate of the state vector:

$$\mathbf{x}_t = (\psi_t, \phi_t, f_t) \tag{3}$$

that defines the homography $\mathtt{H}_t$. Given a certain observation $\mathbf{z}_t$ of the state vector at time step t, the particle filter builds an approxi-

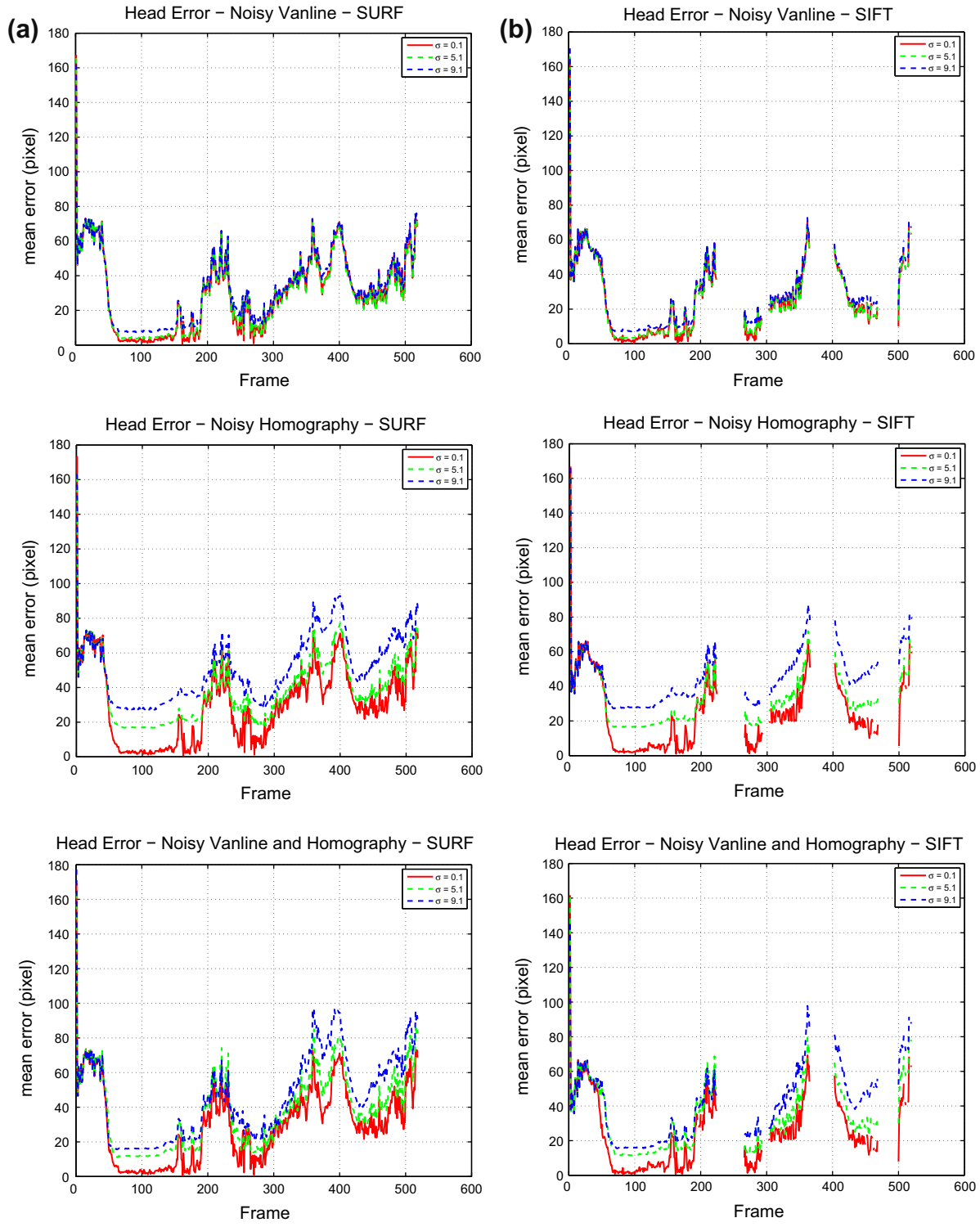

Fig. 8. Head localization accuracy using (a) SURF (b) SIFT keypoint matching for the first sequence (shown in Fig. 7). Top: Errors in vanishing line. Middle: Errors in reference plane to plane homography. Bottom: Errors in vanishing line and in reference plane to plane homography.

mated representation of the posterior pdf $p(\mathbf{x}_t|\mathbf{z}_t)$ through a set of weighted samples $\{(\mathbf{x}_t^i, w_t^i)\}_{i=1}^{N_p}$. Each particle is thus an hypothesis on the state vector value, with a probability associated to it and the estimated value of the state vector is obtained as the weighted sum of all the particles.

The particle filter algorithm requires a probabilistic model for the state evolution and an observation model, from which a prior pdf $p(\mathbf{x}_t|\mathbf{x}_{t-1})$ and a likelihood $p(\mathbf{z}_t|\mathbf{x}_t)$ can be derived. Since there is no prior knowledge about the controls that steer the camera, we adopt a simple random walk model as a state evolution model. This is equivalent to assume that the actual value of the state vector keeps constant through time and relies on the stochastic noise $\mathbf{v}_{t-1}$ to compensate for unmodeled variations, i.e.: $\mathbf{x}_t = \mathbf{x}_{t-1} + \mathbf{v}_{t-1}$, where $\mathbf{v}_{t-1}$ is a zero mean Gaussian process noise with covariance matrix accounting for camera maneuvers.

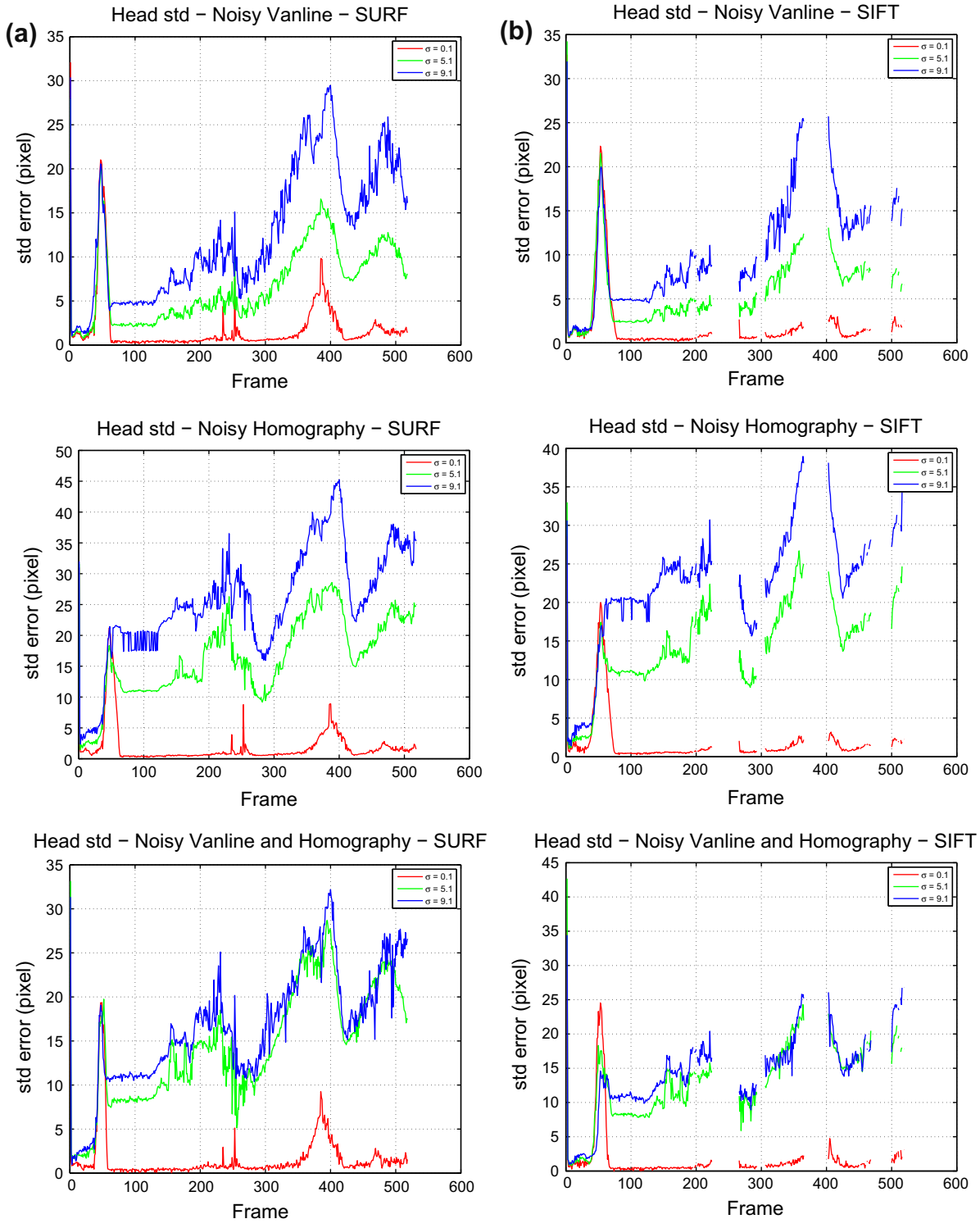

Fig. 9. Head localization standard deviation errors using (a) SURF (b) SIFT keypoint matching for the test sequence #1. Top: STD errors with noisy vanishing line. Middle: STD errors with noisy reference plane to plane homography. Bottom: Both noisy vanishing line and noisy plane to plane homography. Error plot discontinuities in (b) indicate that RANSAC fails to find a consistent homography.

The particle filter uses as observations the correspondences between the SURF keypoints of the current PTZ view and the global map, after that the outliers have been removed. To define the likelihood $p(\mathbf{z}_t|\mathbf{x}_t^i)$ of the observation $\mathbf{z}_t$ generated by the actual camera state given the hypothesis $\mathbf{x}_t^i$ for the PTZ camera parameters we take into account the distance between the backprojections of the corresponding keypoints in $\mathbf{I}_m$ and $\mathbf{I}_t$ in the camera reference plane Π. The keypoint correspondences implicitly suggest the existence of an homography between the camera reference plane and the frame $\mathbf{I}_t$ at time t, and therefore of a triple $(\tilde{\psi}_t, \tilde{\phi}_t, \tilde{f}_t)$ which uniquely describes it. This is performed by estimating the homography $\mathtt{G}_t$ relating $\mathbf{I}_t$ to $\mathbf{I}_m$ using RANSAC. The recovered inliers, the homography $\mathtt{G}_t$ and the homography $\mathtt{H}_{mj}$ associated to the nearest image retrieved $\mathbf{I}_m$ (as registered in the off-line phase, when the scene map was built) are then used to evaluate the likelihood as:

$$p(\mathbf{z}_t|\mathbf{x}_t^i) \propto \exp^{-\frac{1}{\lambda}\sqrt{\sum_{k=1}^{n}\left(\mathtt{H}_t^{i\,-1}\cdot\mathbf{q}_k - \mathtt{H}_{mj}\cdot\mathtt{G}_t\cdot\mathbf{q}_k\right)^2}} \tag{4}$$

where $\mathtt{H}_t^{i\,-1}\cdot\mathbf{q}_k$ and $\mathtt{H}_{mj}\cdot\mathtt{G}_t\cdot\mathbf{q}_k$, $k = 1..n$, are respectively the projection of the predicted and matched keypoints in the camera reference plane Π and λ is a normalization constant. Fig. 5 summarizes this process.

5.2. Sensor slaving: head localization while zooming

Without loss of generality and in order to keep a simple notation, we assume a network with two cameras where $\mathtt{H}_{12}$ is the homography relating the two cameras reference planes. Eq. (1) is now exploited to cooperatively track a target moving in a wide area. According to Eq. (1), the homography $\mathtt{T}_t$ to transfer the imaged target position from the master to the slave camera reduces to:

$$\mathtt{T}_t = \mathtt{H}_t \cdot \mathtt{H}_{12}. \tag{5}$$

Under the assumption of vertical stick-like targets moving on a planar scene the target head can be estimated directly without detecting the target framed by the slave camera. This can be easily done by exploiting the vanishing line in one camera reference plane avoiding the difficulty of detecting the target's head in the images of a moving camera.

For people closely vertical in the scene plane, the position of feet and head can be related by a planar homology [46,12]. According to this, at each time step t, the probability density function of the planar homology $\mathtt{W}_t$ should be computed once the probability density function of respectively the vanishing point $\mathbf{v}_{\infty,t}$ and the vanishing line $\mathbf{l}_{\infty,t}$ in the slave camera view at time t are known.

Once the vanishing line $\mathbf{l}_\infty$ is located in the slave camera reference plane (see Fig. 6), sampling from $p(\mathbf{x}_t|\mathbf{z}_t)$ allows to estimate $p(\mathbf{v}_{\infty,t}|\mathbf{z}_t)$ and $p(\mathbf{l}_{\infty,t}|\mathbf{z}_t)$. For each particle i in the set of the weighted samples $\{(\mathbf{x}_t^i, w_t^i)\}_{i=1}^{N_p}$ that model $\mathtt{H}_t$ we calculate:

$$\mathbf{l}_{\infty,t}^i = [\mathtt{T}_t^i]^{-T}\cdot\mathbf{r}_\infty \tag{6}$$

$$\mathbf{v}_{\infty,t}^i = \omega_t^i \cdot \mathbf{l}_{\infty,t}^i \tag{7}$$

where ω_t^i in Eq. (7) is the dual image of the absolute conic [20] computed as:

$$\omega_t^i = \mathtt{K}_t^i \cdot \mathtt{K}_t^{i\,\mathrm{T}} \tag{8}$$

where the intrinsic camera parameters matrix:

$$\mathtt{K}_t^i = \begin{bmatrix} f_t^i & 0 & p_x \\ 0 & f_t^i & p_y \\ 0 & 0 & 1 \end{bmatrix}$$

is computed with reference to the i-th particle, being f_t^i its estimated focal length component of Eq. (3) and p_x, p_y the coordinates of the principal point located at the image center. From the samples of Eqs. (6)–(8) the pdf $p(\mathtt{W}_t|\mathbf{z}_t) = \frac{1}{N}\sum_{i=1}^{N}\delta(\mathtt{W}_t - \mathtt{W}_t^i)$ is computed as:

$$\mathtt{W}_t^i = \mathtt{I} + (\mu - 1)\,\frac{\mathbf{v}_{\infty,t}^i \cdot \mathbf{l}_{\infty,t}^{i\;\mathrm{T}}}{\mathbf{v}_{\infty,t}^{i\;\mathrm{T}} \cdot \mathbf{l}_{\infty,t}^i}. \tag{9}$$

The cross-ratio μ, being a projective invariant, is the same in any image obtained with the slave camera, while only the vanishing line $\mathbf{l}_{\infty,t}$ and the vanishing point $\mathbf{v}_{\infty,t}$ vary as the camera moves. Thus, the cross-ratio μ can be evaluated accurately by selecting the target feet location $\mathbf{a}$ and the target head location $\mathbf{b}$ in one of the frames, i.e. at time $\bar{t}$ as:

$$\mu = \mathrm{Cross}(\mathbf{v}, \mathbf{a}, \mathbf{b}, \hat{\mathbf{v}}_{\infty,\bar{t}}) \tag{10}$$

where $\mathbf{v}$ is computed as the intersection of the mean vanishing line $\hat{\mathbf{l}}_{\infty,t}$ (averaged over the particles) with the line passing from the mean vanishing point $\hat{\mathbf{v}}_{\infty,t}$ to the feet location $\mathbf{a}$. Using the homogeneous vector representation and the cross product operator, $\mathbf{v}$ can be estimated as: $\mathbf{v} = \hat{\mathbf{l}}_{\infty,\bar{t}} \times (\hat{\mathbf{v}}_{\infty,\bar{t}} \times \mathbf{a})$.

The pdf $p(\mathtt{M}_t|\mathbf{z}_t)$ of the final transformation $\mathtt{M}_t$ that maps the target feet observed in the image of the master camera to the target head in the current image of the slave camera is computed from Eqs. (5) and (9) as:

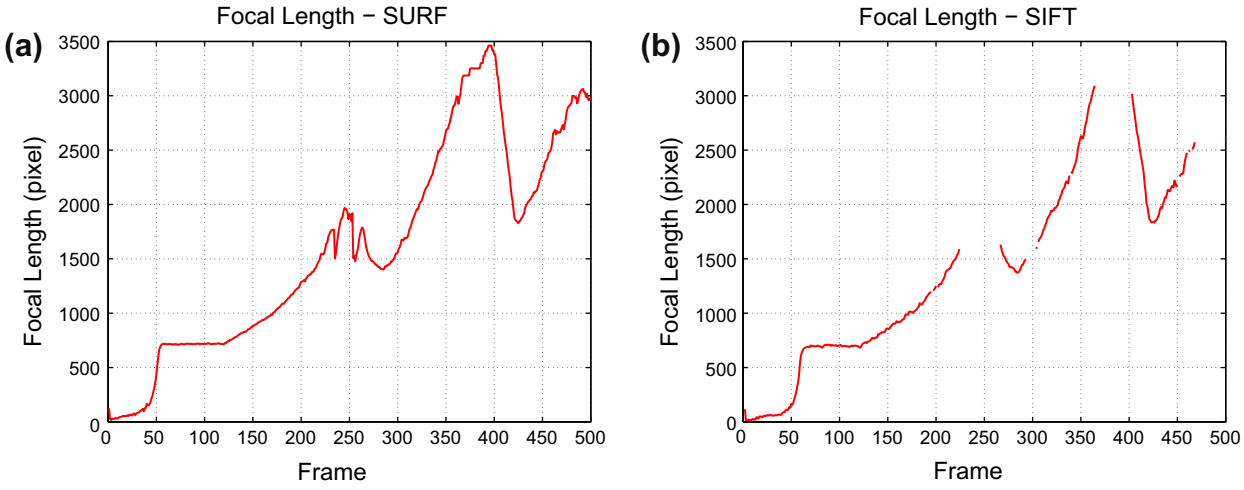

Fig. 10. Estimated camera focal length advancement (in pixel) for the sequence shown in Fig. 7. (a) Using SURF keypoints (b) Using SIFT keypoints. Focal length advancement discontinuities in (b) indicates that RANSAC fails to find a consistent homography.

$$\mathbb{M}_t^i = \mathbb{W}_t^i \cdot \mathbb{T}_t^i = \mathbb{W}_t^i \cdot \mathbb{H}_t^i \cdot \mathbb{H}_{12} \tag{11}$$

where $\mathbb{M}_t^i$ represents a whole family of transformations. Given the estimated $p(\mathbf{x}_t|\mathbf{z}_t)$ of the slave camera and the imaged position of the target as tracked from the master camera, the distribution of the possible head locations $\mathbf{b}_t^i$ as viewed from the slave camera is estimated. We sample L homographies from $p(\mathbf{x}_t|\mathbf{z}_t)$, and the same number of samples from the set of particles tracking the feet position in the master camera view $\mathbf{a}_t^i$, to obtain:

$$\mathbf{b}_t^i = \mathbb{M}_t^i \cdot \mathbf{a}_t^i \qquad i = 1..L \tag{12}$$

It is worth to note that Eq. (12) jointly takes into account both zooming camera calibration uncertainty (through each homography in $\mathbb{M}_t^i$ – see Eq. (11)) and target tracking uncertainty.

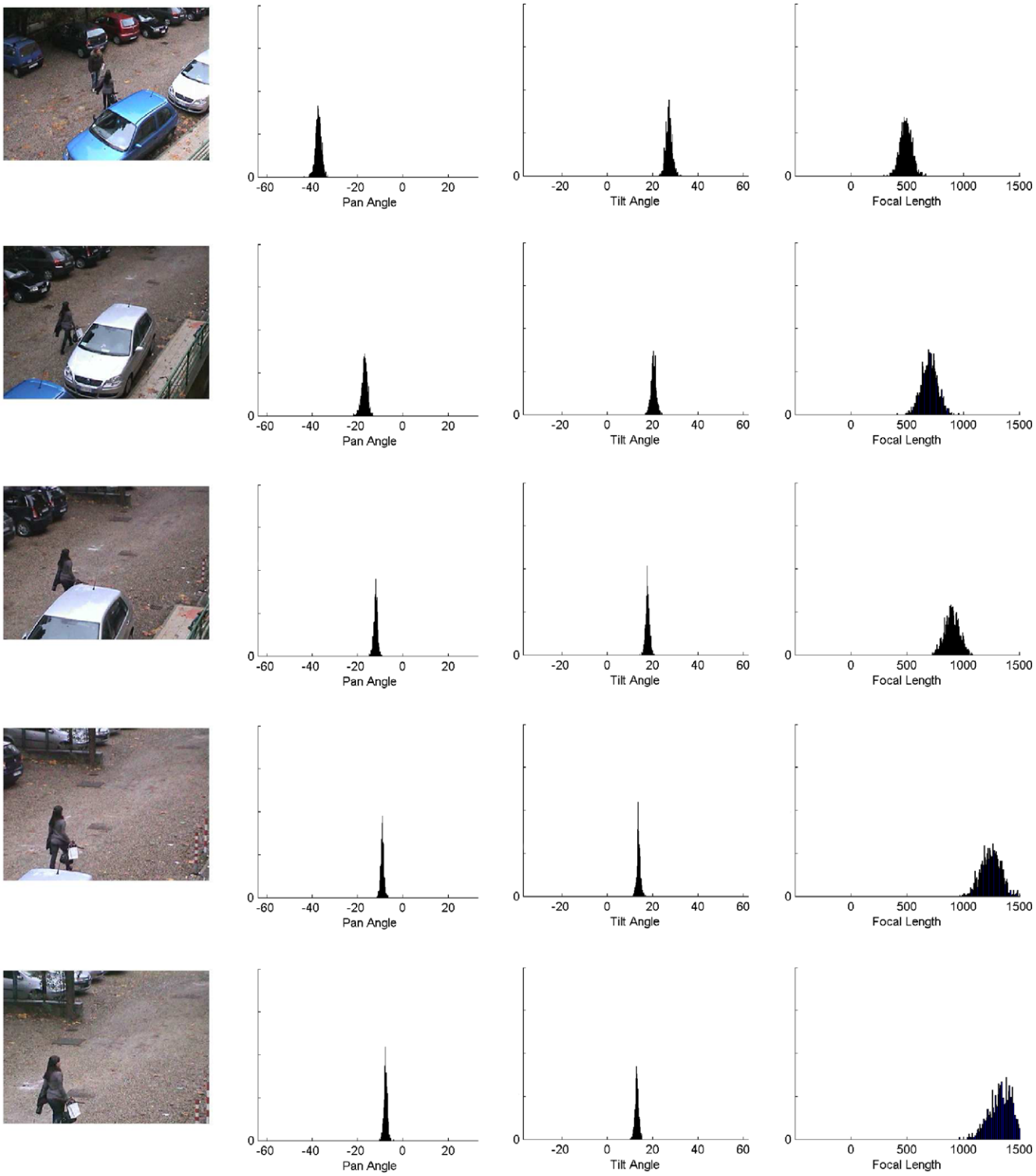

Fig. 11. The pan–tilt–zoom parameters as estimated by the filter for a sample set of frames of the second sequence. The first column shows the sample frames of the sequence; the last three columns show the pan tilt and zoom distributions. In this sequence it is possible to appreciate a high variation for the zoom parameter.

6. Experimental results

The validity of the framework has been evaluated on a special test setup in a wide outdoor parking area of 80x15 meters. Two IP PTZ Sony SNC-RZ30 cameras were placed in proximity of the long side extremes, at about 60 meters far from each other, operating in a master–slave configuration. Images from both cameras were taken at 320×240 pixels of resolution. We used two scene maps of visual landmarks of the observed scene, one from each camera. Each map was built at three zoom factors (wide view, 1.5 times, 2 times) so as to provide a much larger number of feature points at each camera location and to support higher zooming at run-time. The covered field of view is about 180 degrees wide horizontally and 45° wide vertically.

Two different long sequences were used for test. The first was planned to verify the accuracy of the estimation of target's head position in the view of the slave camera that is an important goal in surveillance applications (this permits to drive the camera to appropriately zoom-in on the target's face, for example). In the sequence, the target changes his direction of motion at random, stops and restarts walking. The camera was steered to follow the target, as much as possible, progressively zooming in. After 200 frames the target's head occupies a region of about 20×30 pixels. Because of the sudden changes of motion direction, speed and the trivial camera controller used, a shaking blurred video sequence is obtained. The central position of the head region was manually annotated as the target's head position ground truth for each frame. The second sequence was planned to obtain a qualitative evaluation of the method proposed. In this sequence a person is walking approximatively in a straight path, but the sequence is recorded making panning and zooming continuously in order to follow the trajectory of the target while increasing its scale and keeping it in sight.

In the first test experiment, the accuracy of the method was analyzed with reference to the principal factors influencing the accuracy of target's head estimation: the homography relating the reference planes of the master and the slave camera, the vanishing line in one of the reference planes, the estimated homography $\mathbb{H}_t$ (i.e. the camera parameters and measurements).

Two pairs of parallel lines needed to estimate the vanishing line (see Fig. 6a) and four points needed to estimate the homography relating the two camera reference planes were drawn manually and hence corrupted by a white, zero mean, Gaussian noise with standard deviation between 0.1 and 9 pixel. The influence of this noise over the three factors of influence for the accuracy of head estimation was tested by running a Monte Carlo simulation. The procedure was repeated 1,000 times with different seed for the random noise generator, though with the same noise variance, and averaged over trials. Fig. 7 shows a few frames generated by the method.

Plots of the mean error and standard deviation in head localization as measured when the target feet position in the master camera view is transferred to the head position in the slave camera view are reported respectively in Figs. 8 and 9 for different values of the noise, separately considering: a noisy vanishing line (first row), a noisy master–slave homography (second row) and a noisy vanishing line and master–slave homography (third row). The ef-

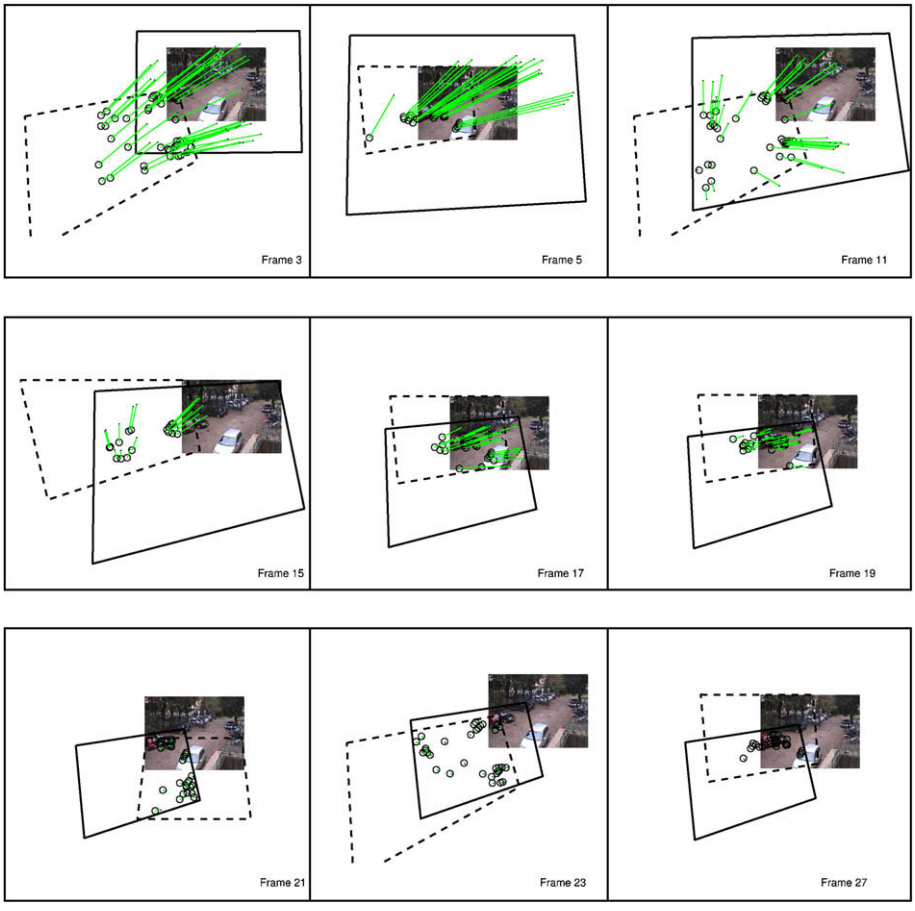

Fig. 12. Filter convergence. The solid polygon indicates the boundary of the current frame $\mathbf{I}_t$ projected onto the reference image plane, through the filtered homography. The dashed polygon indicates the boundary of the matched image $\mathbf{I}_m$ projected onto the reference image plane. The line segments indicate matches (length proportional to the registration error).

fect SURF of keypoints (column (a)) was compared with SIFT (column (b)).

In all the cases, after an initial transient due to the guessed initial conditions of the pan–tilt angles and focal length, the mean error falls to small values and grows almost linearly as the focal length increases. From Fig. 8-(middle row) it appears that head localization is much more sensitive to errors in the master–slave homography rather than the vanishing line. Fig. 9 shows that the standard deviation grows almost proportionally with the noise in the case of errors in the vanishing line, while grows more in the case of noisy homography. Two quick changes of direction of the camera (around frame 250 and 370) to follow the maneuvering target strongly contribute to the uncertainty in target's head localization. In this case it can be appreciated that SURF performs much better than SIFT mostly because it detects more keypoints and makes RANSAC less likely to fail. When the vanishing line and the homography are noisy, both mean errors and standard deviation are smaller than in the other cases. This is due to the fact that the two errors tend to cancel out mainly because they are correlated. Indeed the homography $\mathbb{H}_{lm}$ doubly participates in the final computation of the head coordinates using Eq. (11). The planar homology $\mathbb{W}_t^i$ is in fact parameterized by the vanishing line which is transformed by $\mathbb{H}_{12}$.

The estimated increase of focal length is shown in Fig. 10. It can be observed that error grows also almost linearly as the noise increases. Focal length estimation (that is the most critical factor to achieve head localization accuracy) has a reasonable uncertainty as shown by its time varying distribution. As one might expect, it exhibits a graceful degradation in accuracy under large zoom levels.

Concerning the second experiment, in Fig. 11 are shown the distributions of pan–tilt–zoom camera parameters. From the figure it is possible to observe that the zoom factor has large variations. In Fig. 12 we show a number of frames, including the reference slave camera view (that defines the reference image plane) together with two warped image boundaries superimposed. The solid polygon indicates the boundary of the current frame $\mathbf{I}_t$ projected onto the reference image plane through the estimated homography. The dashed polygon indicates instead the boundary of the matched image $\mathbf{I}_m$ as projected onto the reference image plane. It can be observed that the estimation is initially inaccurate and is hence corrected in a few frames, until the feature points in the current slave view $\mathbf{I}_t$ and the matched feature points in the nearest image $\mathbf{I}_m$ in the map are coincident (circles and dots). Fig. 13 shows an example of the system at work.

The use of a prebuilt map with images taken at multiple zoom levels for each PTZ camera greatly improves the overall performance of the approach with respect to a simpler solution in which a single wide reference view is used, as in [5].

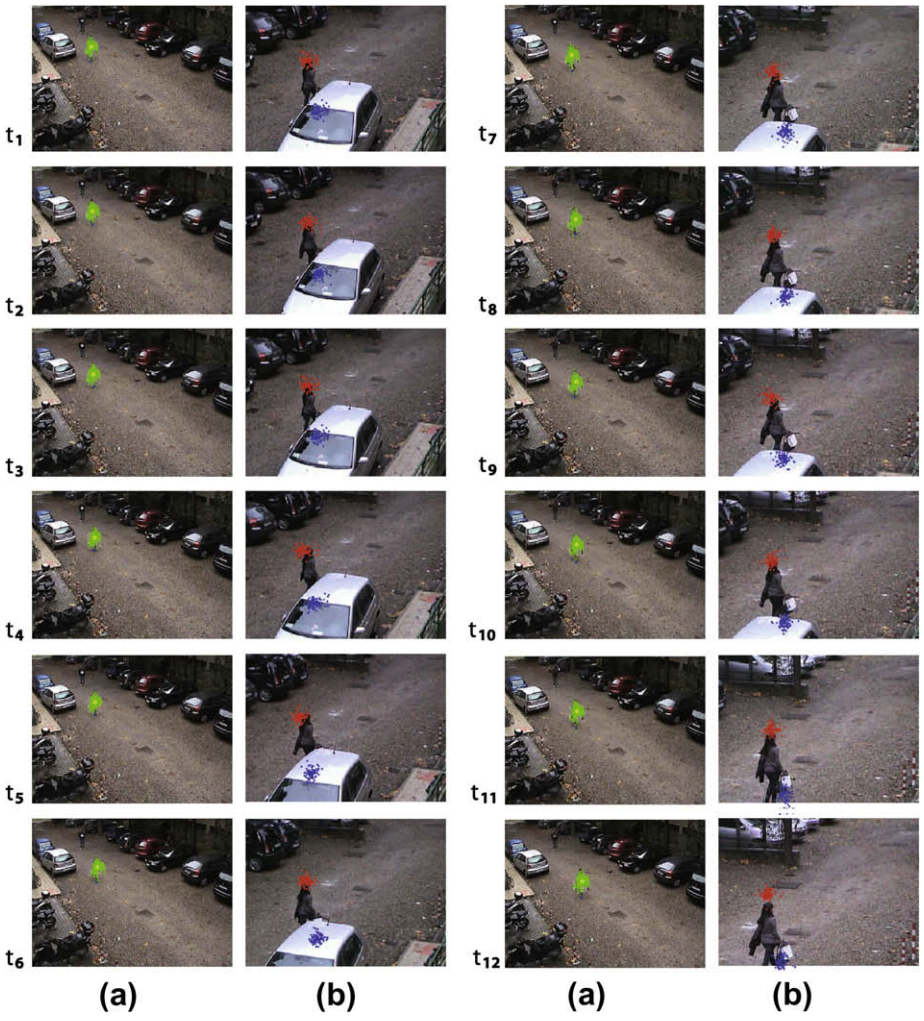

Fig. 13. Twelve frames of the sequence #2 analyzed with the proposed technique. (a) Master camera view: the target is detected by background subtraction. (b) Slave camera view: the particles show the uncertainty of the head and feet position of the target. Although the target is partially occluded the proposed method is still able to localize the target's head.

7. Conclusion

In this paper we have shown how to combine distinctive visual landmarks maps and PTZ camera geometry in order to define and compute the basic building blocks of PTZ camera networks. The proposed approach can be generalized to networks with any arbitrary number of cameras, each of which can act either as master or as slave. The proposed framework does not require any 3D known location to be specified, and allows to take into account both zooming camera and target uncertainties. Results are very encouraging to develop automated biometric identification technologies to identify humans at a distance.

The main limitation of the proposed approach is that the master–slave relationship is estimated using stationary visual landmarks. Instead, most of the available landmarks in the scene are non stationary, especially when observing crowded scenes, or when moving objects determine changes in the scene appearance. Due to this, as time progresses the number of feature matches in the map considerably decreases, until the point at which RANSAC fails to find a consistent homography. Future research will consider the possibility of landmarks maintenance in a continuous changing background.

Acknowledgment

This work is partially supported by the EU IST VidiVideo Project (Contract FP6-045547) and by Thales Italia, Florence, Italy.

References

[1] J. Badri, C. Tilmant, J. Lavest, Q. Pham, P. Sayd, Camera-to-camera mapping for hybrid pan–tilt–zoom sensors calibration, in: SCIA07, 2007, pp. 132–141.
[2] A. Bartoli, N. Dalal, R. Horaud, Motion panoramas, Computer Animation and Virtual Worlds 15 (2004) 501–517.
[3] J. Batista, P. Peixoto, H. Araujo, Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking, in: Proceedings of the IEEE Workshop on Visual Surveillance, 1998, pp. 18–25.
[4] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (surf), Computer Vision and Image Understanding 110 (3) (2008) 346–359.
[5] A.D. Bimbo, F. Dini, A. Grifoni, F. Pernici, Uncalibrated framework for on-line camera cooperation to acquire human head imagery in wide areas, in: IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS'08), 2008.
[6] B. Bose, E. Grimson, Ground plane rectification by tracking moving objects, in: Proceedings of the Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS), October 2003.
[7] D. Chekhlov, M. Pupilli, W. Mayol-Cuevas, A. Calway, Real-time and robust monocular slam using predictive multi-resolution descriptors, in: In 2nd International Symposium on Visual Computing, November 2006.
[8] J. Civera, A.J. Davison, J.A. Magallon, J.M.M. Montiel, Drift-free real-time sequential mosaicing, International Journal of Computer Vision (2008).
[9] Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, 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.
[10] R. Collins, A. Lipton, H. Fujiyoshi, T. Kanade, Algorithms for cooperative multisensor surveillance, Proceedings of the IEEE 89 (10) (2001) 1456–1477.
[11] C.J. Costello, C.P. Diehl, A. Banerjee, H. Fisher, Scheduling an active camera to observe people, in: Proceedings of the 2nd ACM International Workshop on Video Surveillance and Sensor Networks, 2004, pp. 39–45.
[12] A. Criminisi, I. Reid, A. Zisserman, Single view metrology, International Journal of Computer Vision 40 (2) (2000) 123–148.
[13] J. Davis, X. Chen, Calibrating pan–tilt cameras in wide-area surveillance networks, in: Proceedings of ICCV 2003, vol. 1, 2003, pp. 144–150.
[14] A.J. Davison, I.D. Reid, N. Molton, O. Stasse, Monoslam: real-time single camera slam, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (6) (2007) 1052–1067.
[15] L. de Agapito, E. Hayman, I.D. Reid, Self-calibration of rotating and zooming cameras, International Journal of Computer Vision 45 (2) (2001). November.
[16] A. Del Bimbo, F. Dini, A. Grifoni, F. Pernici, Exploiting single view geometry in pan–tilt–zoom camera networks, in: Andrea Cavallaro, Hamid Aghajan (Eds.), Workshop on Multi-camera and Multi-modal Sensor Fusion Algorithms and Applications (M2SFA2 2008), Marseille, France, 2008.
[17] A. del Bimbo, F. Pernici, Distant targets identification as an on-line dynamic vehicle routing problem using an active-zooming camera, in: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS'05) in conjunction with ICCV, Beijing, China, October 2005, pp. 15–21.
[18] A. Hampapur, S. Pankanti, A. Senior, Y.-L. Tian, L. Brown, R. Bolle, Face cataloger: multi-scale imaging for relating identity to location, in: IEEE Conference on Advanced Video and Signal Based Surveillance, 2003, pp. 21–22.
[19] R. Hartley, Self-calibration from multiple views with a rotating camera, in: Proceedings of European Conference on Computer Vision, 1994, pp. 471–478.
[20] R.I. Hartley, A. Zisserman, Multiple View Geometry in Computer Vision, Cambridge University Press, 2000.
[21] R. Horaud, D. Knossow, M. Michaelis, Camera cooperation for achieving visual attention, Machine Vision and Applications 16 (6) (2006) 1–2.
[22] G. Hua, M. Brown, S. Winder, Discriminant embedding for local image descriptors, in: ICCV07, 2007, pp. 1–8.
[23] A. Jain, D. Kopell, K. Kakligian, Y.-F. Wang, Using stationary-dynamic camera assemblies for wide-area video surveillance and selective attention, in: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '06), Washington, DC, USA, 2006, pp. 537–544.
[24] S. Khan, O. Javed, Z. Rasheed, M. Shah, Human tracking in multiple cameras, in: In Proceedings of the International Conference in Computer Vision, vol. I, 2001, pp. 331–336.
[25] D.-W. Kim, K.-S. Hong, Real-time mosaic using sequential graph, Journal of Electronic Imaging 15 (2) (2006) 023005.
[26] G. Klein, D. Murray, Parallel tracking and mapping for small AR workspaces, in: Proceedings of the Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07), Nara, Japan, November 2007.
[27] N. Krahnstoever, P.R.S. Mendonca, Bayesian autocalibration for surveillance, in: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), vol. 2, 2005, pp. 1858–1865.
[28] T. Lee, T. Hollerer, Hybrid feature tracking and user interaction for markerless augmented reality, in: IEEE in Virtual Reality Conference (VR'08), March 2008, pp. 145–152.
[29] S.-N. Lim, D.L.S, A. Elgammal, Scalable image-based multi-camera visual surveillance system, in: Proceedings of IEEE Conference on Advanced Video and Signal Based Surveillance, 2003, pp. 205–212.
[30] Z. Lin, H.-Y. Shum, Fundamental limits of reconstruction-based superresolution algorithms under local translation, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (1) (2004) 83–97.
[31] D.G. Lowe, Distinctive image features from scale-invariant keypoints, International Journal on Computer Vision 60 (2) (2004) 91–110.
[32] F. Lv, T. Zhao, R. Nevatia, Self-calibration of a camera from video of a walking human, in: Proceedings of the 16 th International Conference on Pattern Recognition (ICPR'02), vol. 1, 2002, p. 10562.
[33] M. Grabner, H. Grabner, H. Bischof, Fast approximated sift, in: ACCV06, vol. I, 2006, pp. 918–927.
[34] L. Marchesotti, L. Marcenaro, C. Regazzoni, Dual camera system for face detection in unconstrained environments, in: ICIP, vol. 1, 2003, pp. 681–684.
[35] R. Kumar, A.R. Hanson, Robust methods for estimating pose and a sensitivity analysis, CVGIP 60 (3) (1994) 313–342.
[36] E. Rosten, T. Drummond, Machine learning for high-speed corner detection, in: European Conference on Computer Vision, vol. 1, May 2006, pp. 430–443.
[37] S. Se, D.G. Lowe, J.J. Little, Vision-based mobile robot localization and mapping using scale-invariant features, in: ICRA, 2001, pp. 2051–2058.
[38] A. Senior, A. Hampapur, M. Lu, Acquiring multi-scale images by pan–tilt–zoom control and automatic multi-camera calibration, in: IEEE Workshop on Applications on Computer Vision, 2005.
[39] S. Sinha, M. Pollefeys, Towards calibrating a pan–tilt–zoom cameras network, in: P. Sturm, T. Svoboda, S. Teller (Eds.), OMNIVIS, 2004.
[40] S.N. Sinha, M. Pollefeys, L. McMillan, Camera network calibration from dynamic silhouettes, in: CVPR 2004, 2004.
[41] D. Steedly, C. Pal, R. Szeliski, Efficiently registering video into panoramic mosaics, in: Proceedings of the Tenth IEEE International Conference on Computer Vision, 2005.
[42] S. Stillman, R. Tanawongsuwan, I. Essa, A system for tracking and recognizing multiple people with multiple cameras. Technical report GIT-GVU-98-25 Georgia Institute of Technology, Graphics, Visualization, and Usability Center, 1998.
[43] M.P. Sudipta N Sinha, Jan-Michael Frahm, Y. Genc, Gpu-based video feature tracking and matching, in: Workshop on Edge Computing Using New Commodity Architectures, 2006.
[44] T. Svoboda, H. Hug, L. Van Gool, Viroom – low cost synchronized multi-camera system and its self-calibration, in: Pattern Recognition, 24th DAGM Symposium, Number 2449 in LNCS, September 2002, pp. 515–522.
[45] B. Triggs, P.F. McLauchlan, R.I. Hartley, A.W. Fitzgibbon, Bundle adjustment – a modern synthesis, in: Proceedings of the International Workshop on Vision Algorithms, 2000.
[46] L. Van Gool, M. Proesmans, A. Zisserman, Grouping and invariants using planar homologies, in: Workshop on Geometrical Modeling and Invariants for Computer Vision, Xidian University Press, 1995.
[47] R. Willson, S. Shafer, What is the center of the image?, Journal of the Optical Society of America A 11 (11) (1994) 2946–2955
[48] Q. Zhang, Y. Chen, Y. Zhang, Y. Xu, Sift implementation and optimization for multi-core systems, in: 10th Workshop on Advances in Parallel and Distributed Computational Models (APDCM-08) in conjunction with IPDPS'08, April 2008.
[49] X. Zhou, R. Collins, T. Kanade, P. Metes, A master–slave system to acquire biometric imagery of humans at a distance, in: ACM SIGMM 2003 Workshop on Video Surveillance, 2003, pp. 113–120.