Sensor fusion for cooperative head localization

Alberto Del Bimbo, Fabrizio Dini, Giuseppe Lisanti, Federico Pernici
University of Florence, Italy
Media Integration and Communication Center (MICC)
Florence, Italy
{*delbimbo, dini, lisanti, pernici*}*@dsi.unifi.it*

***Abstract*—In modern video surveillance systems, pan–tilt–zoom (PTZ) cameras certainly have the potential to allow the coverage of wide areas with a much smaller number of sensors, compared to the common approach of fixed camera networks. This paper describes a general framework that aims at exploiting the capabilities of modern PTZ cameras in order to acquire high resolution images of body parts, such as the head, from the observation of pedestrians moving in a wide outdoor area. The framework allows to organize the sensors in a network with arbitrary topology, and to establish pairwise master–slave relationship between them. In this way a slave camera can be steered to acquire imagery of a target keeping into account both target and zooming uncertainties. Experiments show good performance in localizing target's head, independently from the zooming factor of the slave camera.**

***Keywords*-Tracking; Master; Slave; PTZ Camera; Camera Network; Particle Filter;**

I. Introduction

Stationary cameras are not able to monitor a wide area entirely. To overcome this problem, a PTZ camera network can be exploited: several slave PTZ sensors could be controlled by one or more master PTZ camera(s) in order to follow the trajectory of some entities in the scene and generate multi-view close-up imagery at high resolution. In this paper the focus is on establishing at frame-rate the time variant mapping between PTZ cameras present in a network as they redirect the gaze and zoom to acquire high resolution images of moving targets for biometric purpose. The proposed approach 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.At runtime, 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. The use of a prebuilt map with images taken at multiple zoom levels for each PTZ camera improves the performance of the approach with respect to a simpler solution in which a single wide reference view is used, as in [1]. Differently from [2], our solutions explicitly takes into account camera calibration parameters and their uncertainty.

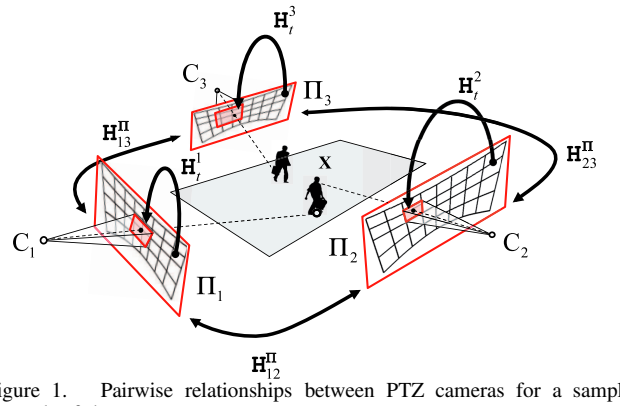

Figure 1. Pairwise relationships between PTZ cameras for a sample network of three cameras.

II. Related work

After the VSAM project [3], new methods have been proposed for calibrating PTZ cameras. Among them, [4] and [5] 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 the feature matching. In [4], a linear mapping is used that is computed from a look-up table of manually established pan and tilt correspondences. In [5], a lookup table is employed that also takes into account camera zooming. In [2], 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.

Real-time estimation of camera's position and orientation using visual landmarks has been proposed by other authors, following the monoSLAM (monocular Simultaneous Localization And Mapping) approach. There, internal camera parameters are typically known in advance [6], while in Structure from Motion techniques they are estimated jointly with 3D structure [7]. Scale-invariant feature transform (SIFT) and matching based on best-bin first K-D tree search [8] were used in [9] for robot localization and mapping to find the visual landmarks and establish their correspondences.

III. Geometric Relationship

Cameras in a network having an overlapping field of view 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 $s \in \{\text{MASTER}, \text{SLAVE}\}$.

As shown in Fig. 1 the three reference planes Π_1, Π_2, Π_3 observed respectively by the cameras $\mathbf{C}_1$, $\mathbf{C}_2$ and $\mathbf{C}_3$ are related to each other through the three homographies $\mathtt{H}_{12}^{\Pi}$, $\mathtt{H}_{13}^{\Pi}$, $\mathtt{H}_{23}^{\Pi}$. Instead, at time t the current image plane is related to the reference plane through the homographies $\mathtt{H}_t^1$, $\mathtt{H}_t^2$ and $\mathtt{H}_t^3$. 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 $\mathtt{H}_{12}^{\Pi}$ and hence from Π_2 to the current zoomed view of $\mathbf{C}_2$ through $\mathtt{H}_t^2$. Referring to the general case of M distinct cameras, once $\mathtt{H}_t^k$ and $\mathtt{H}_{kl}^{\Pi}$, $k \in 1..M$, $l \in 1..M$ 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:

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

IV. Offline learning of the Scene

We consider a pin-hole camera model projecting the three-dimensional world onto a two-dimensional image. Assuming that the camera rotates around its optical center with fixed principal point and without modeling the radial distortion, the projection of a generic image i generated by a camera C, can be 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 [7], 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 ($\theta = 0$). 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 ($f_i \simeq f_j \simeq f$). 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, [10]. Estimates for ψ, ϕ and f can be calculated from the entries of $\mathtt{H}_{ji}$. The image-to-image homography of eq. (2) makes run-time matching and minimization in PTZ camera networks much simpler than with the full 8 DOF homography.

To describe the scene observed by the sensors we collect some images, off-line, at different levels of pan, tilt and zoom, so as to cover the whole field of regard of each PTZ camera. Local distinctive features are used as visual landmarks of the scene to compute the inter-image homographies by exploiting bundle adjustment optimization on the reference images as in [11]. Keypoints and associated camera geometry information are stored in a k-d tree so that quick matching and camera geometry retrieval can be performed in real time.

V. Online Cooperative Head Localization

At runtime keypoints extracted from the current frame $\mathbf{I}_t$ are matched with those in the global k-d tree, according to nearest neighbor search in the feature descriptor space. Once the image $\mathbf{I}_m$ with the highest number of matches is found, the correct homography $\mathtt{G}_t$ relating $\mathbf{I}_t$ to $\mathbf{I}_m$ is computed using RANSAC. The homography $\mathtt{H}_{mj}$ that relates the image $\mathbf{I}_m$ with the image $\mathbf{I}_j$ in the reference plane Π retrieved in the k-d tree is hence used to compute the likelihood to estimate $\mathtt{H}_t$ in eq. (1). To this end, a particle filter is used in order to recover the actual value of the state vector $\mathbf{x}_t = (\psi_t, \phi_t, f_t)$ where ψ_t and ϕ_t are respectively the pan and tilt angle, and f_t is the focal length. This triple completely defines the homography $\mathtt{H}_t$ relating the current frame $\mathbf{I}_t$ to the reference plane.

Given a certain observation $\mathbf{z}_t$ of the state vector at time step t, the particle filter builds an approximated 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. 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. The particle filter uses as observations the correspondences between the SURF keypoints of the current PTZ view and the global map (without outliers). 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 back-projections of the corresponding keypoints in $\mathbf{I}_m$ and $\mathbf{I}_t$ in the camera reference plane Π. 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 $\mathbf{I}_m$ are then used to evaluate the likelihood as:

$$p(\mathbf{z}_t|\mathbf{x}_t^i) \propto \exp^{-\frac{1}{\lambda}\sqrt{\sum_{k=1}^{n}\left(\mathtt{H}_t^{i\,-1}\cdot\mathbf{q}_k - \mathtt{H}_{mj}\cdot\mathtt{G}_t\cdot\mathbf{q}_k\right)^2}} \tag{3}$$

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 the matched keypoints in the camera reference plane Π, while λ is a normalization constant. The entire process is shown in Fig. 2.

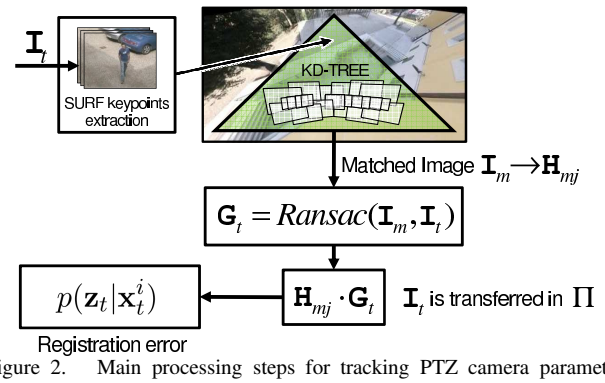

Figure 2. Main processing steps for tracking PTZ camera parameters (online).

Assuming a network of two cameras, where $\mathtt{H}_{12}^{\Pi}$ is the homography relating their reference planes, eq. (1) is reduced to: $\mathtt{T}_t = \mathtt{H}_t \cdot \mathtt{H}_{12}^{\Pi}$. Under the assumption of vertical stick–like targets moving on a planar scene the target head can be estimated directly by a planar homology [12], [13], exploiting the vanishing line in one of the camera reference plane. 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, 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: $\mathbf{l}_{\infty,t}^i = [\mathtt{T}_t^i]^{-\mathrm{T}} \cdot \mathbf{r}_{\infty}$ and $\mathbf{v}_{\infty,t}^i = \omega_t^i \cdot \mathbf{l}_{\infty,t}^i$, where $\mathbf{r}_{\infty}$ is the vanishing line in the master camera view and ω_t^i is the dual image of the absolute conic [7] computed as: $\omega_t^i = \mathtt{K}_t^i \cdot \mathtt{K}_t^{i\mathrm{T}}$. The intrinsic camera parameters matrix $\mathtt{K}_t^i$ is computed with reference to the i-th particle. 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 then computed as:

$$\mathtt{W}_t^i = \mathtt{I} + (\mu - 1)\frac{\mathbf{v}_{\infty,t}^i \cdot \mathbf{l}_{\infty,t}^{i\ \mathrm{T}}}{\mathbf{v}_{\infty,t}^{i\ \mathrm{T}} \cdot \mathbf{l}_{\infty,t}^i}. \tag{4}$$

where μ, namely the cross–ratio, remains the same through the entire sequence, while $\mathbf{l}_{\infty,t}^i$ and $\mathbf{v}_{\infty,t}^i$ are respectively the i^{th} hypothesis for the vanishing line and the vanishing point, varying as the camera moves. 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 as:

$$\mathtt{M}_t^i = \mathtt{W}_t^i \cdot \mathtt{T}_t^i = \mathtt{W}_t^i \cdot \mathtt{H}_t^i \cdot \mathtt{H}_{12}^{\Pi} \tag{5}$$

where $\mathtt{M}_t^i$ 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:

$$\mathbf{b}_t^i = \mathtt{M}_t^i \cdot \mathbf{a}_t^i \qquad i = 1..L \tag{6}$$

It is worth to note that eq. (6) jointly takes into account both zooming camera calibration uncertainty (through each homography in $\mathtt{M}_t^i$ – see eq. (5)) and target tracking uncertainty.

VI. Experimental Results

The tracking accuracy of the proposed framework has been evaluated in a wide outdoor parking area of 80x15 meters, observed by two IP PTZ Sony SNC-RZ30 cameras, working in a master–slave configuration. Images from both cameras were taken at 320×240 pixels of resolution. We build two scene maps (one from each camera) with three different level of zoom factors so as to provide a much larger number of feature points at each camera pose and zoom. Each map takes about 400MB of memory.

To evaluate the head localization error in the slave camera we corrupt the two pairs of parallel lines needed to estimate the vanishing line and the four points needed to estimate the homography $\mathtt{H}_{12}^{\Pi}$, with a white, zero mean, Gaussian noise with standard deviation between 0.1 and 9 pixels. This procedure was repeated 1000 times and averaged over trials. Plots of the mean error in head localization and the estimated increase of focal length are reported in Fig. 3 for different values of the noise. The effect of SURF keypoints (Fig. 3–(a,b)) was compared with SIFT (Fig. 3–(c,d)).

As it can be seen, after a brief transient (necessary to estimate the initial camera pose), the mean error falls to small values and grows almost linearly as the focal length increases. 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. It is also possible to see that errors on the vanishing line and on the homography tend to cancel out, since they are correlated.

Regarding the focal length estimation, it can be seen that the error grows also almost linearly as the noise increases. Some frames (at different levels of zoom) of a sequence analyzed with the proposed technique are shown in Fig. 4. Detailed experimental results can be found in [14].

VII. 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.

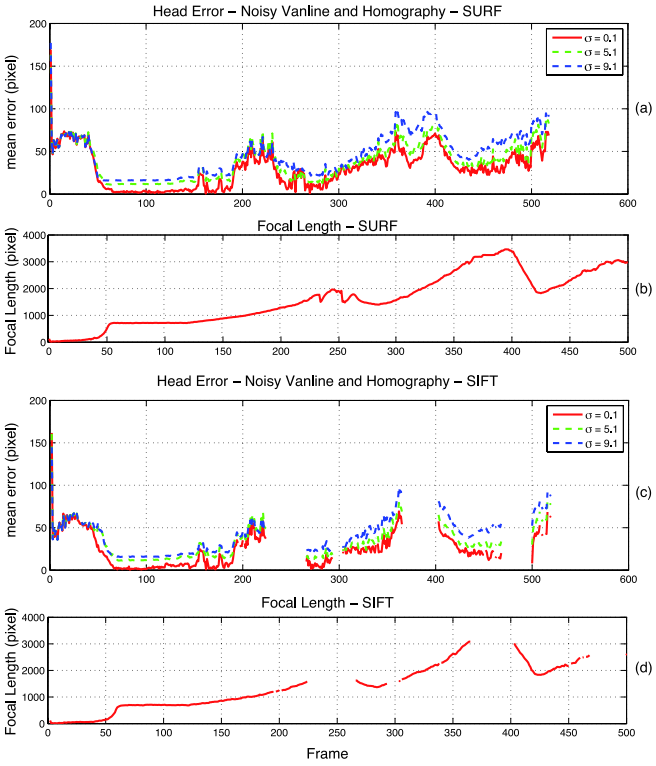

Figure 3. Head localization accuracy and estimated camera focal length advancement (in pixel) using SURF (a,b) and SIFT (c,d) keypoint matching.

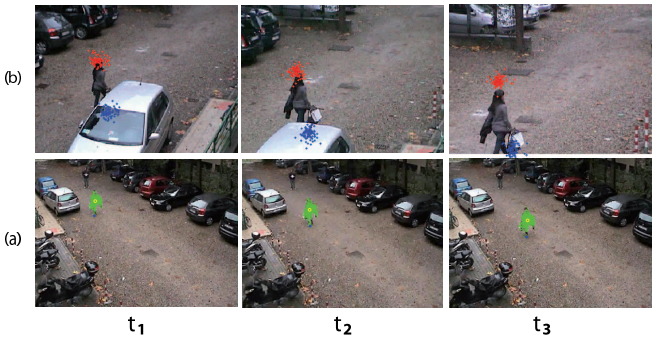

Figure 4. Some frames of a sequence analyzed with the proposed technique. (a) Master camera view: the target is detected by background subtraction. (b) Slave camera view: the particles show the uncertainty of the head and feet position of the target.

The main limitation of the proposed approach is that, as time progresses, the number of feature matches in the map considerably decreases, and this may lead to a failure in the estimation of a consistent homography. Future research will consider the possibility of landmarks maintenance over time in a continuous changing background.

ACKNOWLEDGMENT

This work is partially supported by Thales Italia, Florence, Italy.

REFERENCES

[1] A. Del Bimbo, F. Dini, A. Grifoni, and F. Pernici, "Uncalibrated framework for on-line camera cooperation to acquire human head imagery in wide areas" in *Proc. of IEEE International Conference on Advanced Video and Signal based Surveillance (AVSS)*, Santa Fe, New Mexico, USA, September 2008.

[2] A. Senior, A. Hampapur, and M. Lu, "Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration" *IEEE Workshop on Applications on Computer Vision*, 2005.

[3] Collins, Lipton, Kanade, Fujiyoshi, Duggins, Tsin, Tolliver, Enomoto, and Hasegawa, "A system for video surveillance and monitoring: Vsam final report" *Technical report, Carnegie Mellon University*, May 2000.

[4] 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*, pp. 113–120, 2003.

[5] J. Badri, C. Tilmant, J. Lavest, Q. Pham, and P. Sayd, "Camera-to-camera mapping for hybrid pan-tilt-zoom sensors calibration" in *SCIA07*, 2007, pp. 132–141.

[6] A. J. Davison, I. D. Reid, N. Molton, and O. Stasse., "Monoslam: Real-time single camera slam" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 1052–1067, 2007.

[7] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints" *International Journal on Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] S. Se, D. G. Lowe, and J. J. Little, "Vision-based mobile robot localization and mapping using scale-invariant features" in *ICRA*, 2001, pp. 2051–2058.

[10] A. Bartoli, N. Dalal, and R. Horaud, "Motion panoramas" *Computer Animation and Virtual Worlds*, vol. 15, pp. 501–517, 2004.

[11] L. de Agapito, E. Hayman, and I. D. Reid., "Self-calibration of rotating and zooming cameras" *International Journal of Computer Vision*, vol. 45, no. 2, November 2001.

[12] L. Van Gool, M. Proesmans, and A. Zisserman, "Grouping and invariants using planar homologies" in *Workshop on Geometrical Modeling and Invariants for Computer Vision*. Xidian University Press, 1995.

[13] A. Criminisi, I. Reid, and A. Zisserman, "Single view metrology" *International Journal of Computer Vision*, vol. 40, no. 2, pp. 123–148, November 2000.

[14] A. D. Bimbo, F. Dini, G. Lisanti, and F. Pernici, "Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks" *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 611 – 623, 2010, special Issue on Multi-Camera and Multi-Modal Sensor Fusion.