Camera-to-Camera Mapping for Hybrid Pan-Tilt-Zoom Sensors Calibration

Julie Badri[1,2], Christophe Tilmant[1], Jean-Marc Lavest[1], Quonc-Cong Pham[2], and Patrick Sayd[2]

[1] LASMEA, Blaise Pascal University,
24 avenue des Landais, Aubiere, F-63411 France
[2] CEA, LIST,
Boîte Courrier 65, Gif-sur-Yvette, F-91191 France
{julie.badri,christophe.tilmant,jean-marc.lavest}@univ-bpclermont.fr,
{quoc-cuong.pham,patrick.sayd}@cea.fr

Abstract. Video surveillance becomes more and more extended in industry and often involves automatic calibration system to remain efficient. In this paper, a video-surveillance system that uses stationary-dynamic cameras devices is presented. The static camera is used to monitor a global scene. When it detects a moving object, the Pan-Tilt-Zoom (PTZ) camera is controlled to be centered on this object. We describe a method of camera-to-camera calibration, integrating zoom calibration in order to command the angles and the zoom of the PTZ camera. This method enables to take into account the intrinsic camera parameters, the 3D scene geometry and the fact that the mechanism of inexpensive camera does not fit the classical geometrical model. Finally, some experiment results attest the accuracy of the proposed solution.

Keywords: Camera-to-camera calibration, zoom calibration, visual servoing, Pan-Tilt-Zoom camera, video surveillance.

1 Introduction

Surveillance companies want simultaneously to monitor a wide area with a limited camera network and to record identifiable imagery of all the people passing through that area. To solve this problem, it has been proposed to combine static cameras having a large field-of-view with Pan-Tilt-Zoom cameras. Indeed, it is possible to control the angle of rotation of the PTZ camera (pan and tilt angles) and the zoom. In practice, the system proceeds as follows. A scene event as a moving subject is detected and located using the static camera. The PTZ camera must be controlled with the information of the static camera in order to adjust its pan, tilt and zoom such as the object of interest remains in the field of view. Then, the high resolution image can be recorded in order to apply face or gesture recognition algorithm, for example.

The main problem to solve in this system is how to control the PTZ camera parameters from the information of the object position extracted in the static

B.K. Ersbøll and K.S. Pedersen (Eds.): SCIA 2007, LNCS 4522, pp. 132–141, 2007.

camera. These last years, two approaches emerged. Either, each camera is calibrated in order to obtain the intrinsic and extrinsic camera parameters before to find a general relation between 2D coordinates in the static camera and the pan and tilt angles, like Horaud *et al.* [5] and Jain *et al.* [6] or cameras are not calibrated like Zhou *et al.* [12] and Senior *et al.* [9]. They learned a look-up-table (LUT) linking several positions in the static camera with the corresponding pan-tilt angles. Then, for another point, they estimate the corresponding pan-tilt angles by interpolating using the closest learned values.

In order to position the presented paper, we develop the existing works. In the first case, wthe problem camera calibration, and in particular dynamic camera calibration has been extensively addressed. Either cameras calibration is based on the fact that it is a stereo-vision system like Horaud *et al.* [5] or each camera is calibrated separately like Jain *et al.* [6]. In this case, most existing methods for calibrating a pan-tilt camera suppose simplistic geometry model of motion in which axes of rotation are orthogonal and aligned with the optical center ([1], [2], [10], [4], [11]). If this assumptions can be suitable for expensive mechanisms, they are not to model the true motion of inexpensive pan-tilt mechanisms. In reality a single rotation in pan rotation induces a curved displacement in the image instead of straight lines.

Recently, Jain *et al.* [6] proposed a new calibration method with more degrees of freedom. As with other methods the position and orientation of the camera's axes can be calibrated, but it can be also calibrated the rotation angle. It is more efficient, more accurate and less computationally expensive than the previous works. Actually, Jain *et al.* [6] mean to be the only one to propose a method without simplistic hypothesis. The calibration step involves the presence of a person to deal with the calibration marks. So, this method can not be used in the goal of a turnkey solution for a no-expert public.

Now, methods based on the no-direct camera calibration are focused. Few people have explored this approach. The first were Zhou *et al.* [12] who used collocated cameras whose viewpoints are supposed to be identical. The procedure consisted of collecting a series of pixel location in the stationary camera where a surveillance subject could later appear. For each pixel, the dynamic camera was manually moved to center the image on the subject. The pan and tilt angles were recorded in a LUT indexed by the pixel coordinates in the stationary camera. Intermediate pixels in the stationary camera were obtained by a linear interpolation. At run time, when a subject is located in the stationary camera, the centering maneuver of dynamic camera used the recorded LUT. The advantage of this approach is that calibration marks are not used. This method is based on the 3D information of the scene but the LUT is learned manually.

More recently, Senior *et al.* [9] proposed a procedure more automatic than Zhou *et al.* [12]. To steer the dynamic camera, they need to know a sequence transformations to enable to link a position with the pan-tilt angles. This transformations are adapted to pedestrian tracking. An homography links the foot position of the pedestrian in the static camera with the foot position in the dynamic camera. A transformation links the foot position in the dynamic

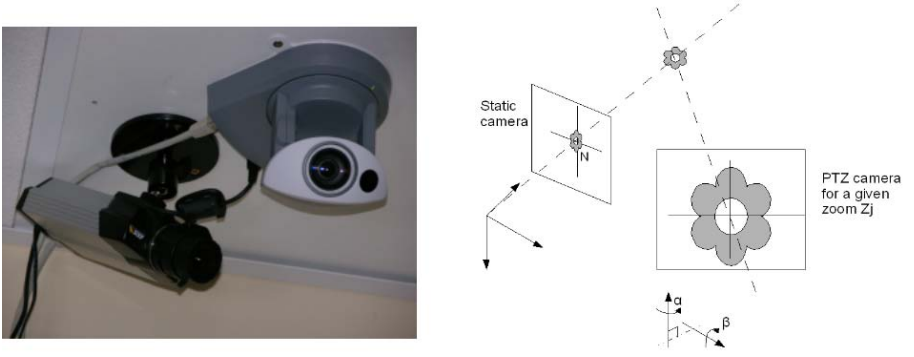

Fig. 1. Our system of collocated cameras : the static camera is on the left and the dynamic camera is on the right

camera with the position in the dynamic camera. Finally, a third transformation encoded in a LUT as Zhou *et al.* [12] links the head position in the dynamic camera with pan-tilt angles. These transformations are learned automatically from unlabelled training data. The default of this method is the step of the establishment of the training data. If this method is used for a turnkey solution for a no-expert public and unfortunately the scene changes, it is impossible that a no-expert public could constitute a good and complete training data in order to update the system.

A solution in the continuity of works of Zhou *et al.* [12] and Senior *et al.* [9] is proposed. Indeed, [6] need the depth information of the object in the scene. So they need to use stereo triangulation. But, like in figure 1, this system is composed of two almost collocated cameras.

Moreover, in the goal of an automatic and autonomous system, the solution of Jain *et al.* [6] and Senior *et al.* [9] are not used. In fact, they need an expert person knowing to use a calibration marks in the case of Jain *et al.* [6] or knowing to extract the good datas to make the training datas in the case of Senior *et al.* [9].

In this paper, an automatic and autonomous solution is presented for a no-calibrated pair of static-dynamic cameras. The solution adapts automatically to its environment. In fact, if the pair of cameras are in a changing environment, this solution can be restarted regularly.

2 Automatic Supervised Multi-sensor Calibration Method

The system used is composed of a static camera with a wide field of view and a Pan-Tilt-Zoom camera with a field-of-view 2.5 times smaller than that of the static camera at the minimal zoom. In the following, the images of the static and the PTZ camera are respectively noted I_s and $I_d(\alpha, \beta, Z)$. The parameters (α, β, Z) represent the pan, tilt and zoom parameters of the PTZ camera.

The method presented in this section enables to learn the relation ζ, for all zoom Z_j, between the pixel coordinates (x_s, y_s) of I_s and the pan-tilt parameters depending of the zoom Z_j :

$$(\alpha_{Z_j}, \beta_{Z_j}) = \zeta(x_s, y_s, Z_j) \; . \tag{1}$$

To learn the relation ζ, two steps are needed. The first step is the computation of an camera-to-camera mapping (LUT). The LUT gives the relation between n_s pre-defined pixels of I_s and the pan-tilt parameters such as the pixel is mapped to the center C_d of $I_d(\alpha, \beta, Z)$ for different samples of the zoom $Z_{j=0,10,...,m}$. In the following, the n_s pre-defined pixels of I_s are called nodes and noted $\mathbf{N} = \left\{N_s^0, N_s^1, ..., N_s^{n_s-1}\right\}$. The second step is the extension of the LUT for all the pixel of I_s and all values of the zoom Z.

2.1 Camera-to-Camera Calibration : 3D Scene Constraints Integration in LUT Computation

Computation of the LUT. The computation of the LUT integrates two loops: (1) computation of the LUT for a constant zoom Z_0 for all the nodes of $\mathbf{N}$ and (2) computation of the LUT for each zoom $Z_{j=0,1,...,m}$ for all the nodes of $\mathbf{N}$.

To begin the computation of the LUT at the zoom Z_0, we need to be in the neighbourhood V_0 of N_s^0. In order to move automatically the PTZ camera in V_0, pan-tilt parameters are randomly selected until the field-of-view of the PTZ is in a good neighbourhood of N_s^0.

The main steps of this procedure are :

1. Initialization on N_s^0;
2. For **each node** N_s^i in the static camera :
   (a) Selection of images I_s and $I_d(\alpha, \beta, Z_0)$ to be compared
   (b) Extraction and robust matching of interest points between I'_s and $I_d(\alpha, \beta, Z_0)$
   (c) Computation of an homography H between interest points of I'_s and $I_d(\alpha, \beta, Z_0)$
   (d) Computation of the $N_{i,s}$ coordinates in $I_d(\alpha, \beta, Z_0)$: $N_d^i = H \times N_s^i$
   (e) Command of the dynamic camera in order to N_d^i catch up with C_d
   (f) Process $N_{i,s}$ until the condition $|N_{i,d} - C_d| < \epsilon$ is reached. Otherwise we stop the loop after k loops
3. Go to the step (2) to process the node N_s^{i+1};

At the step (2a), a small image I'_s is extracted from the complete image I_s around the node N_s^i to process in order to optimize the matching result. In fact, the field-of-view of the PTZ camera is smaller than the one of the static camera. So, the size of I'_s is defined such as the field-of-view of I'_s is nearly the same that the field-of-view of the PTZ camera.

For the step (2b), the scale-invariant feature transform (SIFT) method proposed by Lowe [7] for extracting and matching distinctive features from images of static and PTZ cameras is used. The features are invariant to image scale,

rotation, and partially invariant to changing viewpoints, and change in illumination.

At the step (2c), we assume that locally the area in the static and PTZ cameras can be approximated by a plane. Locally, the distortion in I_s can be considered insignificant. So, the homography H is searched such as it is the homography that best matches a set of points extracts of the lists of points obtained previously. The set of correspondences contains a lot of outliers. So, the homography H is robustly estimated with a RANSAC procedure. A homography is randomly computed from only four points, and test, how many other points satisfy it many times. The optimal solution is the homography which has the highest number of good points.

When the coordinates of N_s^i in I_d are known, the parameters (α, β) of the PTZ camera must be estimated in order to insure the convergence of N_d^i to the center C_d. We use a proportional controller based on the error between the coordinates of N_d^i and the coordinates of C_d, such as it minimizes the criterion of the step (2e). We assume that the pan-tilt axes and the coordinates axes are collocated. So we can write :

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} K_{x\to\alpha} & 0 \\ 0 & K_{y\to\beta} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \tag{2}$$

The error $(\Delta x, \Delta y)$ corresponds to the coordinates of $N_{i,d} - C_d$. As we are in a static training problem, a proportional controller is sufficient.

This procedure is repeated as long as $|N_d^i - C_d| < \epsilon$ is not achieved. If the system diverges, we stop after k loops.

After convergence, to steer the PTZ camera in the neighbourhood of the next point N_s^{i+1}, the needed angles are estimated with the knowledge of the previous learned points. For a new point N_s^{i+1}, we search the closest point among the previous processed points for each direction. The pan-tilt parameters of the best result are used to move the PTZ camera in the neighbourhood of N_s^{i+1}.

For the computation of the LUT for the other zooms, the same procedure is used. For a given zoom Z_j, instead of comparing a small image of I_s with an image I_d for a node N_s^i, an image I_d at the zoom Z_{j-1} centerred on the node N_s^i is used as the reference image for the visual servoing of the PTZ camera on the node N_s^i at the zoom Z_j.

Construction of N. The choice of the n_s nodes N_s^i depends on the information contained in the 3D scene. For an image I_s, interest points are extracted with SIFT method. For each pixel of I_s, we compute the density function of the interest points with the Parzen window method. The size of the window depends on the relation between each field-of-view of cameras. Then, we search the pixel in I_s with the maximal probability : it is the first node N_s^0 of $\mathbf{N}$. The value zero is given to the probability of the pixels around N_s^0 in order to obtain a best repartition of nodes. This procedure is repeated until the last pixel of I_s with a no-zero probability. The figure 2 shows the nodes obtained with this procedure.

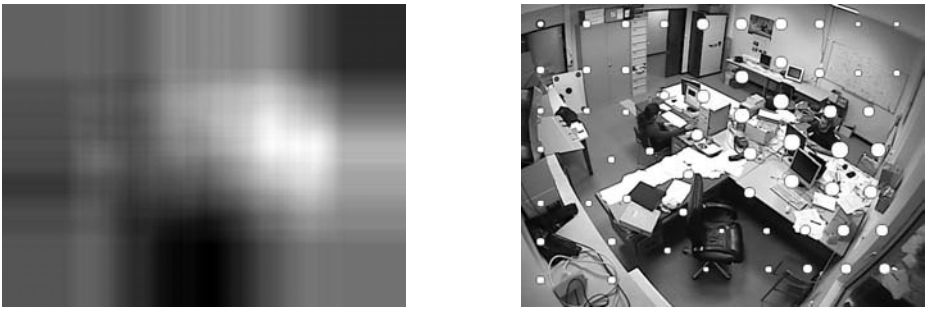

Fig. 2. The left figure represents the estimation of the density function of the interest points in I_s by using the Parzen window method. The right figure shows an example of the extracted grid to I_s, more the circle is important more the density function is important.

2.2 Expansion of LUT

After the previous section 2.1, we obtain a LUT for n_s pixels of the static image I_s for m different zooms. In order to complete the LUT for all the pixels of I_s and all values of the zoom, an approximation of this data is searched.

In such interpolating problems, Thin-Plate-Spline (TPS) interpolation, proposed by Bookstein *et al.* [3], is often preferred to polynomial interpolation because it gives similar results, even when using low degree polynomials and avoids Runge's phenomenon for higher degrees (oscillation between the interpolate points with a big variation). A TPS is a special function defined piecewise by polynomials.

The computation of the interpolation relation ζ resulting of the TPS method needs a training step. The correspondences between the coordinates of a pixel of I_s for a given zoom Z_j and the pan-tilt parameters for Z_j learned during the

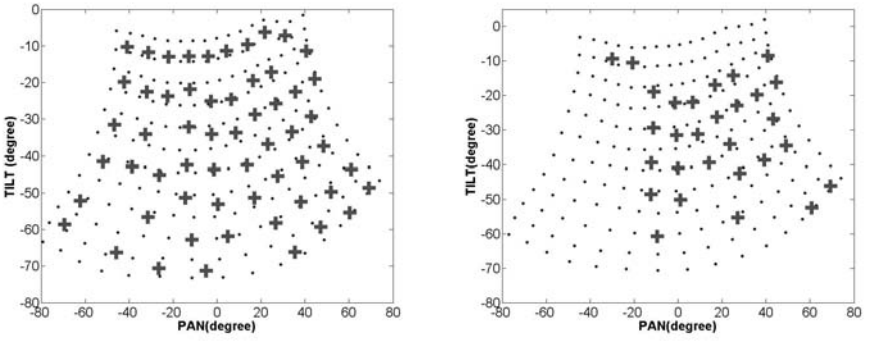

Fig. 3. Result of the TPS interpolation method : zoom Z_0 on the left figure and zoom Z_9 on the right figure. Plus correspond to the results of the computation of the LUT for the nodes of $\mathbf{N}$. Points correspond to the use of ζ for unknown pixels of I_s.

LUT computation is used as training data for TPS method. So, for all triplet (x_s, y_s, Z_j), the pan-tilt parameters can be estimate with ζ, see figure 3.

3 Results and Experiments

Cameras of the AXIS company are used. The image resolution used is 640×480 pixels for the static camera and 704×576 pixels for the PTZ camera. The field of view of the static camera is around $90\,^\circ$. In the case of minimal zoom, the field of view of the PTZ camera is of $42\,^\circ$.

The PTZ camera has a 26x optical zoom. The difference of the field of view between the two cameras can be important. For example, the field of view of the PTZ camera is 2.5 times smaller than one of the static camera at the zoom Z_0, 5 times smaller at the zoom Z_4 and 12.5 times at the zoom Z_7.

The mechanical step of the PTZ camera is $0.11\,^\circ$. Experimentally, we show that the mean displacement in the image I_d for the minimal mechanical step depends on the zoom, see table 1. At best, the accuracy of the solution is limited by this mechanical factor.

Table 1. Mean displacement in pixel in the image of the PTZ camera for different zooms

Zoom	0	1	2	3	4	5	6	7	8
mean in X (pixels)	1.64	2.03	2.42	2.59	3.75	4.76	6.93	9.47	15.71
mean in Y (pixels)	1.57	2.99	2.48	3.54	5.03	5.43	7.92	10.46	13.97

In order to estimate the accuracy of this supervised calibration method, it is necessary to know exactly the coordinates of a reference pixel P_s in I_s and to find precisely its coordinates in I_d. To solve this problem, a black ellipsis E which is visible in the two cases is used. To determine with accuracy the coordinates of the center of E, the binarization method of Otsu [8] is used. Pixels of a region of interest can be separated in two classes. Then, the coordinates of the center of gravity of black pixels are estimated with subpixelic precision.

3.1 Accuracy of the Visual Servoing Loop

For evaluating the accuracy of the visual servoing loop (see section 2.1) at the zoom Z_0, three positions of the ellipsis E are choosen : (1) E on a node issued of the Parzen window method with a high range, (2) a node with a middle range and (3) a node with a small range.

The result of the experimentation is given on the figure 4 with the absolute errors of pan and tilt. For the case 1, the standard deviation is small and of the same order of the mechanical step ($0.11\,^\circ$). For the other two cases, the number of interest points around the nodes is less important so the error and the standard deviation increase. The estimation of the pan-tilt parameters is less accuracy. This result shows that it is important to choose an to classify the nodes in function of the density of interest points around the nodes.

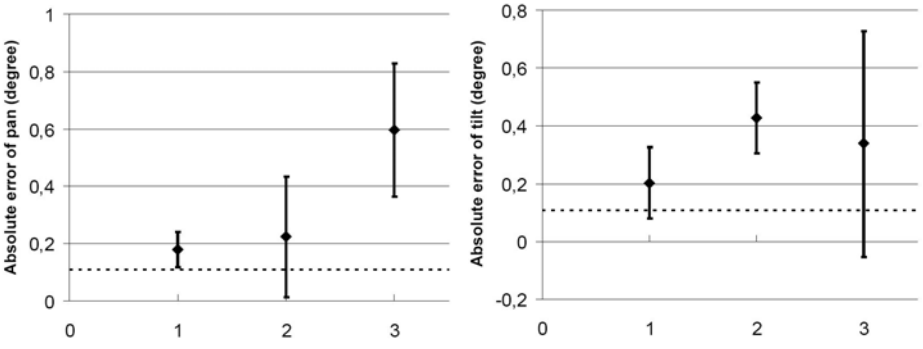

Fig. 4. Absolute errors of pan and tilt in degree resulting of the visual servoing loop for three positions n the scene. The dot line represents the mechanical step.

3.2 Accuracy of the Supervised Calibration Method

For evaluate the accuracy of the complete solution (see section 2), three cases are choosen : (1) E on a pixel which is on the same 3D plane that the closer learned nodes, (2) E on a pixel which is on a different 3D plane that each closer learned node and (3) E on an unknow object when ζ is learned, see figure 5.

The result of the experimentation is given on the figure 6 with the error normalized to mechanical step for the parameters pan-tilt. The case 1 (triangle) is the most ideal case because the 3D information is homogeneous. So, the error is small. The case 2 (diamond) is more complex because the 3D information of the scene presents big variations. So, the error is bigger than the case 1. In the case 3 (square), the unknown object modify the geometry of the scene. This variation was not learned during the learning step of ζ. So the error is more important than the previous cases. Moreover, we note on the figure 6 the error increases with the zoom. Along the learning of the LUT for several zoom, the

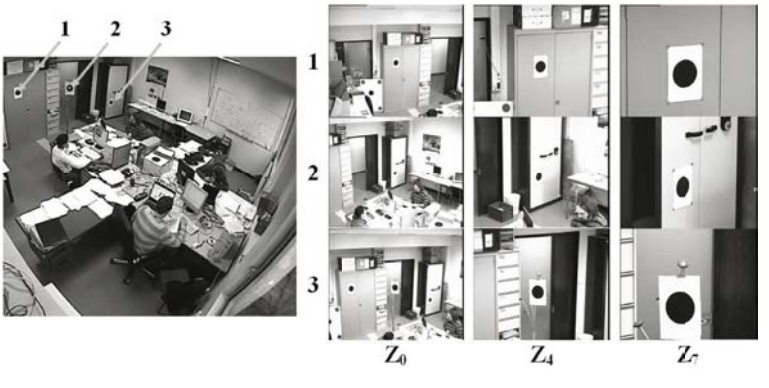

Fig. 5. Illustration of the method. The left figure represents the static camera with the three positions of the target noted. On the right figure, the result for the PTZ camera is shown for different levels of zoom.

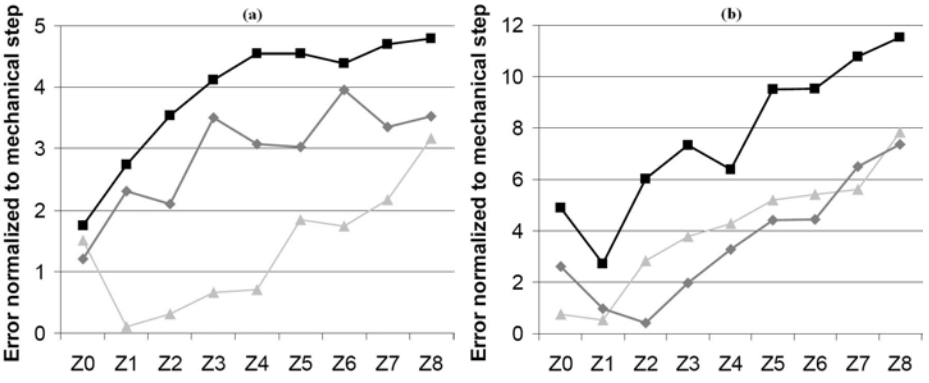

Fig. 6. Error normalized to mechanical step for pan (a) and tilt (b) parameters for several zoom for three cases in the 3D scene (see figure 5) : triangle for case 1, diamond for case 2 and square for case 3

comparison is made between two consecutive zooms. So, the error accumulates progressively. But, the figure 5 shows that even for high zoom, the result can be a good initialization to track a person.

4 Conclusion and Perspectives

In this paper, an automatic algorithm of camera-to-camera calibration integrating the zoom calibration was presented in order to steer a PTZ camera using information of the static camera. At the end, we obtain the relation ζ, for all zoom Z_j, between the pixel coordinates (x_s, y_s) of I_s and the pan-tilt parameters depending on the zoom Z_j : $(\alpha_{Z_j}, \beta_{Z_j}) = \zeta(x_s, y_s, Z_j)$.

All the parameters are automatically selected. The process includes a measure of grey level activity (SIFT). We want to apply the system with an automatic reconfiguration to develop a tracking survey system to focus on the human face along a sequence of displacement.

In the future, in order to reduce the error resulting from the step of the LUT learning for several levels of zoom, the envisaged solution is to compare the zoom Z_j with the zoom Z_0 and to change the zoom Z_0 for a high zoom when the difference between the zoom Z_j and Z_0 is too important. Then, after this amelioration, the solution will be tested in real condition of people tracking.

References

1. Barreto, J.P., Peixoto, P., Batista, J., Araujo, H.: Tracking Multiple Objects in 3D. IEEE Intelligent Robots and Systems. IEEE Computer Society Press, Los Alamitos (1999)
2. Basu, A., Ravi, K.: Active camera calibration using pan, tilt and roll. IEEE Transactions on Systems Man and Cybernetics. IEEE Computer Society Press, Los Alamitos (1997)

3. Bookstein, F.L.: Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence. IEEE Computer Society Press, Los Alamitos (1989)
4. Davis, J., Chen, X.: Calibrating pan-tilt cameras in wide-area surveillance networks. ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision, p. 144. IEEE Computer Society, Washington (2003)
5. Horaud, R., Knossow, D., Michaelis, M.: Camera cooperation for achieving visual attention. Machine Vision Application, vol. 16, pp. 1–2. Springer, Heidelberg (2006)
6. Jain, A., Kopell, D., Kakligian, K., Wang, Y.-F.: Using Stationary-Dynamic Camera Assemblies for Wide-area Video Surveillance and Selective Attention. IEEE Computer Vision and Pattern Recognition, pp. 537–544. IEEE Computer Society, Los Alamitos (2006)
7. Lowe, D.G.: Object Recognition from Local Scale-Invariant Features. ICCV '99: Proceedings of the International Conference on Computer Vision, vol. 2, p. 1150. IEEE Computer Society, Washington (1999)
8. Otsu, N.: A threshold selection method from grey scale histogram. IEEE Transactions on Systems Man and Cybernetics, vol. 1, pp. 62–66. IEEE Computer Society Press, Los Alamitos (1979)
9. Senior, A.W., Hampapur, A., Lu, M.: Acquiring Multi-Scale Images by Pan-Tilt-Zoom Control and Automatic Multi-Camera Calibration. WACV-MOTION '05: Proceedings of the Seventh IEEE Workshops on Application of Computer Vision, vol. 1, pp. 433–438. IEEE Computer Society, Washington (2005)
10. Shih, S., Hung, Y., Lin, W.: Calibration of an active binocular head. IEEE Transactions on Systems Man and Cybernetics. IEEE Computer Society Press, Los Alamitos (1998)
11. Woo, D.C., Capson, D.W.: 3D visual tracking using a network of low-cost pan/tilt cameras. canadian Conference on Electrical and Computer Engineering Conference Proceedings. IEEE Computer Society Press, Los Alamitos (2000)
12. Zhou, X., Collins, R.T., Kanade, T., Metes, P.: A master-slave system to acquire biometric imagery of humans at distance. IWVS '03: First ACM SIGMM international workshop on Video surveillance, pp. 113–120. ACM Press, New York (2003)