5

# Automatic Calibration of Hybrid Dynamic Vision System for High Resolution Object Tracking

Julie Badri[1,2], Christophe Tilmant[1], Jean-Marc Lavest[1],
Patrick Sayd[2] and Quoc Cuong Pham[2]
[1]*LASMEA, Blaise Pascal University, 24 avenue des Landais, Aubiere, F-63411*
[2] *CEA LIST, Boîte Courrier 94, Gif-sur-Yvette, F-91191*
*France*

## 1. Introduction

Visual object recognition and tracking require a good resolution of the object to accurately model its appearance. In addition, tracking systems must be able to robustly recover moving target trajectory, and possibly cope with fast motion and large displacements. Wide angle static cameras capture a global view of the scene but they suffer from a lack of resolution in the case of a large distance between the objects and the sensor. On the contrary, dynamic sensors such as Pan-Tilt-Zoom (PTZ) cameras are controlled to focus on a 3D point in the scene and give access to high resolution images by adapting their zoom level. However, when a PTZ camera focuses on a target, its very limited field of view makes the tracking difficult. To overcome these limitations, hybrid sensor systems composed of a wide angle static camera and a dynamic camera can be used. Coupling these two types of sensors enables the exploitation of their complementary desired properties while limiting their respective drawbacks.
Calibration is required to enable information exchange between the two sensors to produce collaborative algorithms. The calibration of our system is difficult because of changes of both intrinsic (focal length, central point, distortion) and extrinsic (position, orientation) parameters of the dynamic sensor during system exploitation. Two approaches for dynamic stereo sensor calibration are possible:

- **Strong calibration** involves a complete modeling of the system. Intrinsic parameters of each camera and extrinsic parameters are estimated. This approach enables the projection of 3D points, expressed in the world frame, in 2D points expressed in each image frame.
- **Weak calibration** does not target to estimate intrinsic or extrinsic parameters. The objective is only to estimate the direct relation between pixels of the different sensors. From a pixel in the first camera, which is the projection of a given 3D point, the calibration gives the projection of the same 3D point to the second camera. In this approach, the recovery of 3D point coordinates is not more difficult.

### 1.1 The strong calibration approach

Our system is composed of two cameras observing the same scene (see Fig. 1).
We denote:



Source: Pattern Recognition Techniques, Technology and Applications, Book edited by: Peng-Yeng Yin,
ISBN 978-953-7619-24-4, pp. 626, November 2008, I-Tech, Vienna, Austria

- $P_w$ a 3D point of the scene. The 3D coordinates of $P_w$ are expressed in the world reference frame $R_w$.
- $P_{I_s}$ the projection of $P_w$ in the image $I_s$ from the static sensor. The 2D coordinates of $P_{I_s}$ are expressed in the image frame $R_{I_s}$
- $P_{I_d}$ the projection of $P_w$ in the image $I_d$ from the dynamic sensor. The 2D coordinates of $P_{I_d}$ are expressed in the image frame $R_{I_d}$.

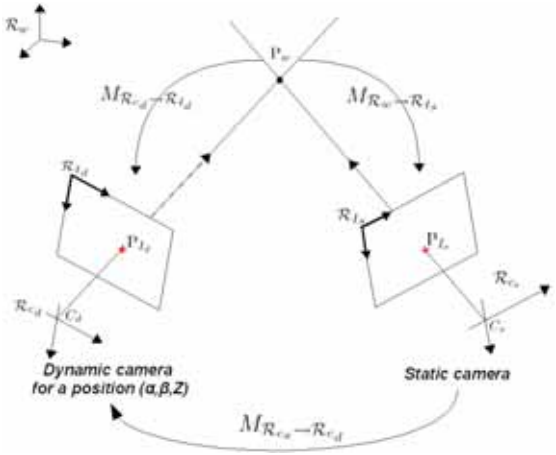


Fig. 1. Vision system and geometric relations.

The strong calibration enables the computation of the coordinates of $P_{I_s}$ and $P_{I_d}$ from the 3D coordinates of $P_w$. Reciprocally, the 3D coordinates of $P_w$ can be inferred from the coordinates of $P_{I_s}$ and $P_{I_d}$ (triangulation). The calibration process consists in estimating the transformation matrices $M_{R_w \to R_{I_s}}$ from the world frame $R_w$ to the image frame $R_{I_s}$ of the static sensor and $M_{R_w \to R_{I_d}}$ from the world frame $R_w$ to the image frame $R_{I_d}$ of the dynamic sensor.

In the next section, we present methods for the calibration of the static sensor. Then, we present the calibration of the dynamic sensor following the same objectives as the static sensor calibration, but with its specific constraint. The third section is dedicated to solutions to gather the two sensors in the same world frame.

**Static camera calibration**

The pin-hole camera model is a usually used to represent image formation for standard camera. This model supposes that all light rays converge through a point $C_s$ called the optical center (see Fig. 2). The focal length $f$ represents the distance from the optical center to the image plane. The optical axis is defined by the point $C_s$ and is orthogonal to the image plane. The principal point $O$ is defined as the intersection of the optical axis with the image plane. The 3D world is projected on the image plane following a perspective transformation.

The calibration of the static sensor consists in estimating the transformation matrix $M_{R_w \to R_{I_s}}$ which is composed of the extrinsic transformation $M_{R_w \to R_{C_s}}$ from the world frame to the camera frame $R_{C_s}$ and the intrinsic transformation from the camera frame $R_{C_s}$ to the image frame $R_{I_s}$ .

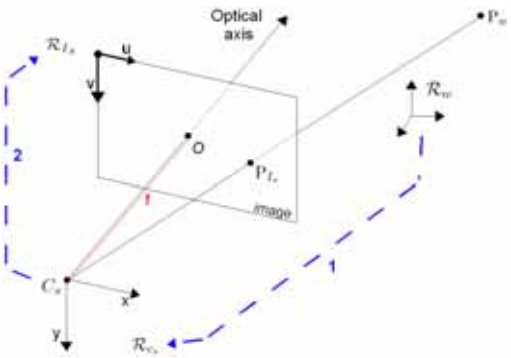


Fig. 2. The pinhole camera model for image formation.

$M_{R_w \to R_{C_s}}$ is composed of a rotation, denoted $R$, and a translation, denoted $t$. The transformation between $P_w$ and $P_{cs}$ is described as follows:

$$P_{c_s} = M_{R_w \to R_{C_s}} P_w = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \\ 0 & 0 & 0 & 1 \end{pmatrix} P_w = \begin{pmatrix} R & t \\ 0^t & 1 \end{pmatrix} P_w \tag{1}$$

As proposed in (Horaud & Monga, 1995), the transformation $M_{R_{Cs} \to R_{Is}}$ is expressed as follows:

$$P_{I_s} = M_{R_{C_s} \to R_{I_s}} P_w = \begin{pmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{pmatrix} P_{C_s} \tag{2}$$

It supposes that rows and columns of the sensors are orthogonal. The parameters $k_u$ and $k_v$ are respectively horizontal and vertical scale factors (expressed in pixels per length unit), $u_0$ and $v_0$ are the coordinates of the principal point $O$. These parameters are called intrinsic parameters.

$$\begin{aligned} P_{I_s} &= M_{R_w \to R_{I_s}} P_w \\ P_{I_s} &= \begin{pmatrix} k_u f & 0 & u_0 & 0 \\ 0 & k_v f & v_0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} R & t \\ 0^t & 1 \end{pmatrix} P_w \end{aligned} \tag{3}$$

Furthermore, optical distortions must be taken into account with real cameras. With standard cameras, two parameters are required to model radial and tangential distortion by polynomial functions (Lavest & Rives, 2003).
A standard approach to solve equation (3) consists in estimating both intrinsic and extrinsic camera parameters from the position of a known pattern (see Fig. 3). The first step of the calibration procedure deals with the accurate detection of pattern features on the calibration pattern. Several types of features can be used: cross (Peuchot, 1994), center of ellipse (Lavest et al., 1998; Brand & Mohr, 1994) and other techniques (Blaszka & Deriche, 1995). The extracted features serve as inputs of a non linear optimization process where the criterion to minimize is generally the sum of quadratic errors measured between the pattern features and their re-projection using the estimated camera model.

Fig. 3. Examples of calibration patterns.

**Dynamic camera calibration**
Geometric calibration of a dynamic sensor is much more complex than a standard camera one. Indeed, most of the proposed methods suppose a simple cinematic model (Fig. 4) where the rotation axes are orthogonal and centered on the optical axis (Barreto et al., 1999; Basu & Ravi, 1997; Collins & Tsin, 1999; Fry et al., 2000; Horaud et al., 2006; Woo & Capson, 2000).
Under this assumption, the camera geometric model can be represented by the following equation:

$$P_{I_d} = M_{R_{C_d} \to R_{I_d}} R_y R_x M_{R_w \to R_{C_d}} P_w \tag{4}$$

where $R_x$ represents the pan rotation matrix and $R_y$ the tilt rotation matrix.

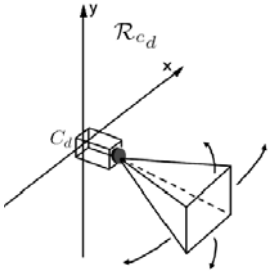


Fig. 4. Simplified cinematic model of a dynamic camera (Davis & Chen, 2003). The rotation axes (pan and tilt) are centered on the optical axis.

Standard dynamic cameras do not respect the constraint of rotation axes centered on the optical center because it is not compliant with low cost mechanic production. Indeed,

rotation mechanisms are independent for pan and tilt. Furthermore, there is the motion of the optical center during zoom changes which makes this assumption unrealistic.
The modeling of standard mechanisms requires the introduction of an additive degree of freedom in the command equation (5). Davis and Chen (Davis & Chen, 2003) proposed a general formulation for this equation, which was extended later in (Jain et al., 2006).

$$\begin{aligned} P_{I_d} &= M_{R_{C_d}\to R_{I_d}} t_y^{-1} R_y t_y t_x^{-1} R_x t_x M_{R_w\to R_{C_d}} P_w \\ P_{I_d} &= M_{R_w\to R_{I_d}}(\Lambda, x, y) P_w \end{aligned} \tag{5}$$

where $R_x$ (resp. $R_y$) represents the rotation matrix in pan (resp. tilt), $t_x$ (resp. $t_y$) represents the horizontal (resp. vertical) translation of the optical center. $\Lambda$ represents the intrinsic and extrinsic parameters of the dynamic camera.

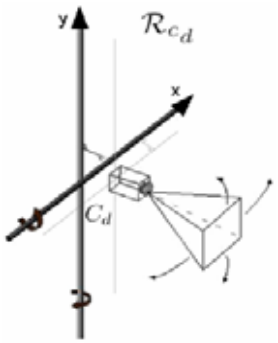


Fig. 5. Generalized cinematic model of the dynamic camera (Davis & Chen, 2003). Pan and tilt motions are represented by arrows and modeled as a rotation around a random 3D direction.

To determine $M_{R_w\to R_{I_d}}(\Lambda, x, y)$, a finite set of angle pairs $(\alpha_i, \beta_i)$ is regularly sampled in the range of the dynamic sensor motion. For each couple $(\alpha_i, \beta_i)$, the dynamic camera is considered as a static camera and it is calibrated with standard techniques. A set of correspondences between 3D points and their 2D projection in the image is built. Camera parameters are estimated by minimizing the differences between the projection of 3D points and their associated observations. Instead of using a passive calibration pattern, the authors use an active pattern composed of LEDs (Light-Emitting Diodes) in order to cover the complete field of view of the dynamic camera.
In (Jain et al., 2006), in addition to the calibration of rotation axes in position and orientation, correspondences between the camera command angles and the real observed angles are searched for. The extended method includes the following complementary steps:

1. Construction with interpolation of the expressions $\hat{\alpha} = g(\alpha)$ and $\hat{\beta} = g(\beta)$, which link the required angles $\alpha$ and $\beta$ with the real ones $\hat{\alpha}$ and $\hat{\beta}$.
2. Construction by interpolation of transformations $t_x$ and $t_y$ with respect to zoom: for a given number of zoom values, the relative position of the optical center and the rotation axis are recorded.

**A common reference frame**

When the static and the dynamic cameras are calibrated, a common reference frame definition is required. As shown in Fig. 1, the following equations can be derived:

$$\begin{aligned} P_{C_s} &= M_{R_w \to R_{C_s}} P_w \\ P_{C_d} &= M_{R_w \to R_{C_d}} P_w \\ P_{C_d} &= M_{R_{C_s} \to R_{C_d}} P_{C_s} \end{aligned} \tag{6}$$

The three transformations are dependent:

$$\begin{aligned} M_{R_w \to R_{C_s}} &= M^{-1}{}_{R_{C_s} \to R_{C_d}} M_{R_w \to R_{C_d}} \\ M_{R_w \to R_{C_d}} &= M_{R_{C_s} \to R_{C_d}} M_{R_w \to R_{C_s}} \\ M_{R_{C_s} \to R_{C_d}} &= M_{R_w \to R_{C_s}} M^{-1}{}_{R_w \to R_{C_d}} \end{aligned} \tag{7}$$

The matrix $M_{R_{C_s} \to R_{C_d}}$ is easily determined from the calibration of each sensor, and can be written:

$$M_{R_{C_s} \to R_{C_d}} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & b_x \\ r_{21} & r_{22} & r_{23} & b_y \\ r_{31} & r_{32} & r_{33} & b_z \\ 0 & 0 & 0 & 1 \end{pmatrix} \tag{8}$$

where the vector **b** = $\begin{pmatrix} b_x & b_y & b_z \end{pmatrix}^t$ is the translation between the optical centers of the static camera and the dynamic camera.

**Conclusion on strong calibration**

Strong calibration gives a complete geometric modeling of the pair of sensors. Knowing the projection model for each sensor and the spatial relation between the sensors, coordinates of 3D points can be inferred from their observations in images. This property is fundamental to recover 3D information for reconstruction purpose. Large orientation angles between the two sensors reduce the uncertainty on 3D reconstruction even if it complicates data matching between images.

These methods are based on the use of calibration patterns, and require human intervention. This constraint is not compatible with the objective to obtain a system able to adapt itself to environment changes, implying automatic re-calibration.

### 1.2 Weak calibration of the dynamic stereo sensor

In many computer vision applications such as object tracking and recognition, a pair of cameras with close points of view, make visual information matching possible (see Fig. 6). However, in this case, it was shown that the estimation of motion parameters become difficult, particularly for small angles in the dynamic sensor (Gardel, 2004). Weak calibration solves this problem, because it enables the estimation of the dynamic camera command from visual information extracted in static images, without analytic modeling of the vision system. The basic idea is to find a mapping between pixels coordinates in the static camera

and rotation angles of the dynamic sensor, at a given zoom value. Moreover, weak calibration avoids explicit modeling of optical distortions. This approach also implicitly encodes the 3D structure of the observed scene.

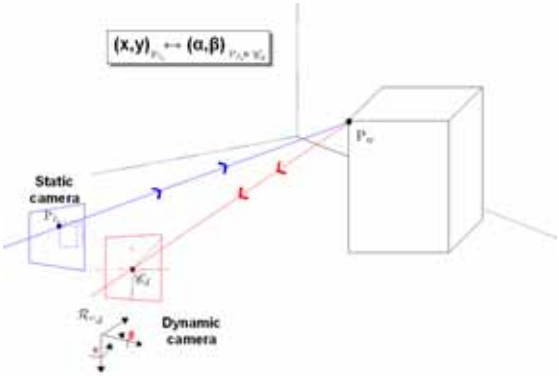


Fig. 6. Weak calibration of a pair of sensors.

Zhou et al. (Zhou et al., 2003) proposed an implementation of this method. A *lookup table* (LUT) linking a pixel of the static camera with the pan and tilt angles centering the dynamic sensor on the corresponding 3D point is built. The LUT is created in two steps:

- **Creation of an LUT for a set of points $P_{I_s}$ in the static sensor**: for each point $P_{I_s}^k$, the dynamic sensor is manually commanded to set the center of the dynamic image on the corresponding point $P_w^k$ of the real scene. The $P_{I_s}^k$ coordinates and pan-tilt angles $(\alpha_i, \beta_i)$ are recorded in the LUT.
- **Interpolation for all the pixels of the static sensor:** a linear interpolation is done between the pixels of the initial set. This linear interpolation is not adapted to handling optical distortions and nonlinear 3D geometry variations. The precision is acceptable to initialize a person tracking and so set up the PTZ camera, as the object is in the field of view, but not to perform pixel matching for intensive sensor collaboration. A denser initial set of points could lead to better accuracy, but it would require a considerable amount of intervention of the supervisor to control the PTZ.

More recently, Senior et al. (Senior et al., 2005) presented a calibration system applied to people tracking where the slave camera is steered to a pan/tilt position calculated using a sequence of transformations, as shown in Fig. 7. Each transformation is learned from unlabelled training data, generated by synchronized video tracking of people in each camera. The method is based on the assumption that people move on a plane and a homography is sufficient to map ground plane points (the location of the feet) in the master camera into points in the second camera. The homography $H$ is learned using the approach described in (Stauffer & Tieu, 2003), and the transformation $T$ inferred from the learned mapping between pan-tilt angles $(\alpha, \beta)$ generated on a spiral and the motion of the optical center in the dynamic camera compared to the known home position $(x_0, y_0)$ where the camera correspondence homography was trained. Then, $T$ is estimated by solving a least-squares linear system $\Theta = TX$ where $\Theta$ represents all couples $(\alpha, \beta)$, and $X$ all coordinates $(x_i - x_{0,} y_i - y_0)$ corresponding to $(\alpha, \beta)$.

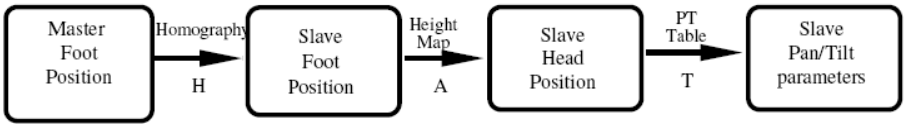


Fig. 7. Calibration approach proposed in (Senior et al., 2005): sequence of transformations to control the PTZ camera PTZ using tracking results in the static image.

### 1.3 Discussion on the choice of the calibration method for our vision system

Our objective is to develop information fusion between the two sensors. We chose to set cameras close to each other to facilitate image matching. This option led us to consider weak calibration (excluding 3D triangulation possibility). This choice is reinforced by material consideration. The low-cost PTZ camera makes strong calibration approach difficult (focal length management).

Weak calibration methods (Zhou et al., 2003; Senior et al., 2005) are manual or require expert skill contribution to elaborate learning bases. These constraints are not compatible with an autonomous and self calibrating system. We propose in the following a weak calibration method that requires no human intervention. Our contribution concerns the two main objectives:

- **Automatism and autonomy:** the proposed method is based on the construction of an LUT which associates pixels of the static sensor with pan-tilt angles to center the dynamic sensor on the corresponding scene point. Our approach exploits natural information, without using a calibration pattern or any supervised learning base. This automatic approach makes re-calibration possible during the system's life and thus drastically reduces the requirement on human intervention.
- **Precision:** the approach uses an interpolation function to get a correspondence for all pixels of the static image. This approach also takes into account distortions in images.

## 2. Learning-based calibration of the hybrid dynamic sensor system

### 2.1 Overview of the system

The hybrid dynamic vision system is composed of a static wide angle camera and a dynamic (Pan-Tilt-Zoom) camera. In the following, the images of the static and the PTZ cameras are respectively denoted $I_s$ and $I_d(\alpha,\beta,Z)$. The parameters $(\alpha,\beta,Z)$ represent the pan, tilt, and zoom parameters of the PTZ camera.

The proposed calibration method can be considered as a registration process by visual servoing. It consists in learning the mapping $\zeta$, for any zoom level $Z$, between the pixel coordinates $(x_s,y_s)$ of a point $P_{I_s}$ of the static camera and the pan-tilt command angles $(\alpha_Z,\beta_Z)$ to be applied to center the dynamic camera on the corresponding point $P_{I_d}$:

$$(\alpha_Z,\beta_Z)=\zeta(x_s,y_s,Z) \tag{9}$$

The data registration relies on the extraction of interest points in regions of interest, which are visually matched in the two images. The basic assumption for interest point matching is that there is locally enough texture information in the image. Moreover, in order to speed up the calibration procedure, the mapping between the two cameras is not computed for all pixels

$P_{I_s}$ in the static camera. Thus, correspondences are searched for in a subset of pixels $\Gamma\{P_{I_s}\}$. The complete mapping is then estimated by interpolation and coded in the LUT.
The learning of the mapping $\zeta$ is performed in two main steps:

1. Automatic sub area registration of the two cameras views for a subset of pre-defined positions $\Gamma\{P_{I_s}\}$ by visual servoing, at different zoom values $Z_{j=0,1,\cdots,m}$ (see Fig. 8)
   a. Learn the mapping $\zeta$ at the minimum zoom level, denoted $Z_0$, for pixels in $\Gamma\{P_{I_s}\}$
   b. Learn the mapping $\zeta$ at sampled zoom values $Z_{j=0,1,\cdots,m}$, for pixels in $\Gamma\{P_{I_s}\}$.
2. Automatic global area matching by interpolation for all pixels of $I_S$ and all values of the zoom.

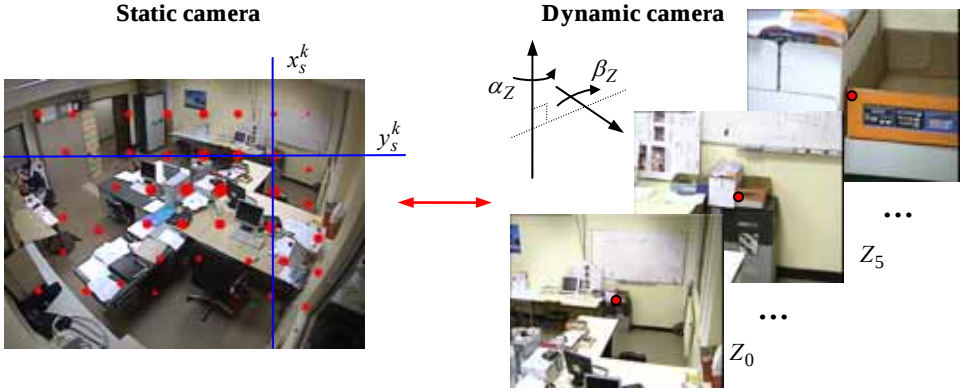


Fig. **8**. Learning of the mapping $\zeta$ between a pixel $(x_s, y_s)$ in the static camera and the pan-tilt angles $(\alpha_Z, \beta_Z)$ to be applied in the PTZ camera, at a given zoom level $Z$ by visual servoing. The learning is performed for a subset of pre-defined points.

### 2.2 Calibrating the hybrid dynamic sensor system at $Z_0$

The proposed calibration method can be compared to the *Iterative Closest Point* (ICP) algorithm. The ICP was first presented by Chen and Medioni (Chen & Medioni, 1991) and Besl and McKay (Besl & McKay, 1992). This simple algorithm iteratively registers two points sets by finding the best rigid transform between the two datasets in the least squares sense.
In our calibration approach, points sets are registered such that the angular parameters of the PTZ camera are optimal, e.g. the point $P_{I_d}$ corresponding to a considered point $P_{I_s}$ is moved to the center $C_d$ of the dynamic camera image.
The algorithm for registering the camera sub areas for points in $\Gamma\{P_{I_s}\}$ at $Z_0$ is summarized below:

1. Start with point $P_{I_s}^0$
2. For each point $P_{I_s}^k$ of the selected points subset $\Gamma\{P_{I_s}\}$
   a. Selection of images $I_s$ and $I_d(\alpha, \beta, Z)$ to be compared

b. Detection and robust matching of interest points between a region of interest $I'_s$ of $I_s$ around $P_{I_s}^k$ and $I_d(\alpha,\beta,Z)$

c. Estimation of the homography $H$ between interest points of $I'_s$ and $I_d(\alpha,\beta,Z)$

d. Computation of $P_{I_d}^k$ coordinates in $I_d(\alpha,\beta,Z)$ by $P_{I_d}^k = H \times P_{I_s}^k$

e. Command of the dynamic camera in order to $P_{I_d}^k$ catch up with $C_d$

f. Process $P_{I_s}^k$ until the condition $\left|P_{I_d}^k - C_d\right| < \varepsilon$ is reached. Otherwise, the algorithm stops after $n$ iterations.

3. Go to step (2) to process the next point $P_{I_s}^{k+1}$.

The main difficulty for registering images from a hybrid camera system resides in the heterogeneity of image resolutions and a potentially variable visual appearance of objects in the two sensors in terms of contrast and color levels for instance.

The registration procedure thus requires a method for detecting and matching visual features robust to scale, rotation, viewpoint, and lightning. In (Mikolajczyk & Schmidt, 2005), the performance of state-of-the art feature matching methods is evaluated. The *Scale-Invariant Feature Transform* (SIFT) (Lowe, 1999) exhibits great performance regarding these constraints.

Because the field of view of the dynamic camera is smaller than that of the static camera, interest points are detected in a region of interest $I'_s$ of $I_s$ around the point $P_{I_s}^k$ such that $I'_s$ approximately corresponds to the view of the dynamic camera. The estimated registration error is taken as the distance in pixels between $P_{I_d}$ and $C_d$. Consequently, we must be able to calculate the coordinates of $P_{I_d}$ and the transform between a point in the static camera and its corresponding point in the dynamic camera. We make the assumption that interest points in $I'_s$ and $I_d(\alpha,\beta,Z)$ are linked by a homography $H$, which means that interest points are supposed to locally lie in a plane in the 3-D scene. Moreover, the distortion in the static camera is considered locally insignificant. The homography $H$ is robustly computed with a RANSAC algorithm (Fischler & Bolles, 1981).

In order to ensure the convergence of $P_{I_d}$ to $C_d$, we use a proportional controller based on the error between the coordinates of $P_{I_d}$ and the coordinates of $C_d$ so that it minimizes the criterion of step (2.f). Assuming that the pan-tilt axes and the coordinate axes are collocated for small displacements, we can write:

$$\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} K_{x\to\alpha} & 0 \\ 0 & K_{x\to\beta} \end{pmatrix} \begin{pmatrix} \Delta x \\ \Delta y \end{pmatrix} \tag{10}$$

$$\text{where } (\Delta x \quad \Delta y)^T = P_{I_d} - C_d$$

During the learning stage, the 3-D scene is assumed to be invariant. As the calibration procedure is an off-line process, there is no temporal constraint on the speed of the PTZ

command. As system accuracy depends mainly on the mechanics of the dynamic camera, a proportional controller is sufficient for calibration purpose.

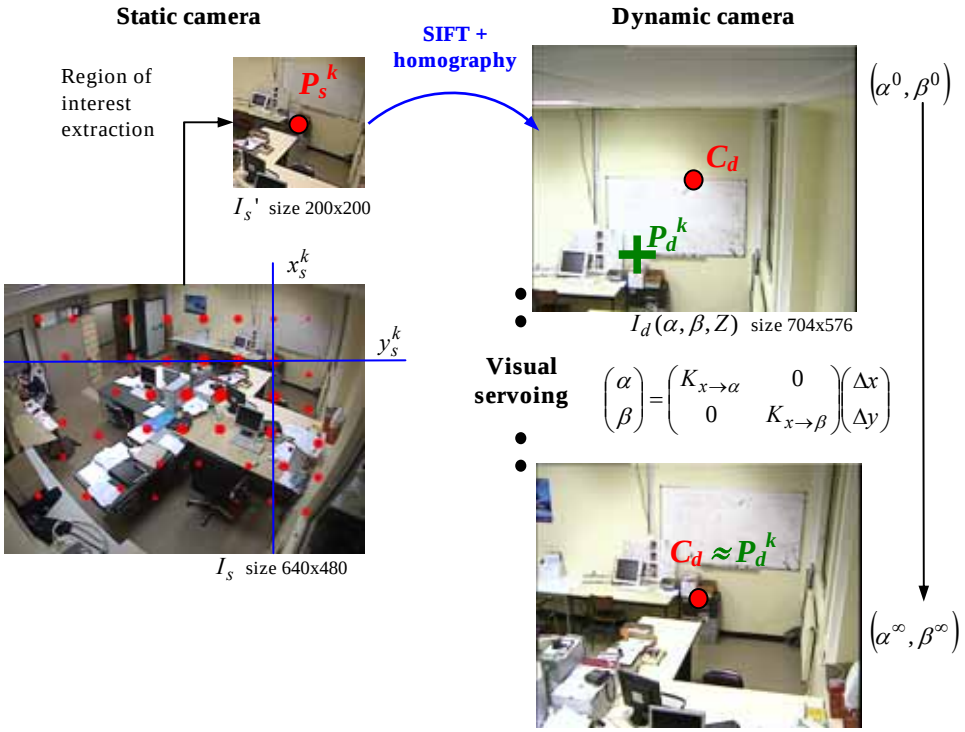


Fig. 9. Calibration of the hybrid sensor system at zoom $Z_0$ by visual servoing on a subset of pre-defined points.

### 2.3 Initialization of the dynamic camera for a given learning point $P_{I_s}$

**Case of the initial point** $P_{I_s}^0$

The objective of the initialization stage is to obtain a pair of pan-tilt angles such that the projection of $P_{I_s}^0$ and its neighborhood in the dynamic camera are visible in the image $I_d$. Starting from the obtained configuration, the dynamic camera parameters are optimized through the visual servoing iterations. The initialization algorithm is given here.

Let $P_{I_s}^{C_d}$ be the projection of $C_d$ in $I_s$.

Repeat until $P_{I_s}^{C_d}$ falls in the neighborhood of $P_{I_s}^0$:

1. Random generation of a pair of $(\alpha,\beta)$ values
2. Selection of current images $I_s$ and $I_d(\alpha,\beta,Z)$ to be compared
3. Detection and robust matching of interest points between $I_s$ and $I_d(\alpha,\beta,Z)$

4. Computation of homography $H$ between matched interest points in $I_s$ and $I_d(\alpha, \beta, Z)$
5. Computation of $C_d$ coordinates in $I_s$ by $P_{I_s}^{C_d} = H \times C_d$

**Case of pre-defined $P_{I_s}^k$ points**

Let us assume that $m$ grid points $\left\{P_{I_s}^k\right\}_{k=1,\cdots,m}$ are already learned. To move the dynamic camera in the neighborhood of $P_{I_s}^{m+1}$, we estimate the command parameters $(\alpha^{m+1}, \beta^{m+1})$ from previously learned positions. For each independent direction, the closest point to $P_{I_s}^{m+1}$ is searched for in the learning base, and serves as an initialization for the calibration on the current point.

### 2.4 Calibrating the hybrid dynamic sensor system at other zoom values

The presented learning algorithm for the system calibration at the initial zoom $Z_0$ and for a subset of pre-defined grid points can be applied at other zoom levels. The only difference is the selection of the reference image. At zoom $Z_0$, the image from the static camera is compared to the image in the dynamic camera. For other zoom values, the images to be registered come from the dynamic sensor at two different zoom levels $Z_k < Z_j$.

Instead of taking $k = j - 1$, which would incrementally cause an accumulation of calibration errors, we select $Z_k$ as the minimum zoom value so that the two images can be registered.
The main steps of the method are:

1. For each point $P_{I_s}^k$
   a. Initialize the dynamic camera at a reference zoom $Z_{ref}$ (initially set to $Z_0$) using the previous learned command $(\alpha, \beta)$
   b. Select the current image in the dynamic camera, denoted $I_d^{ref}$
   c. For each zoom value $Z_j$ so that $Z_j < Z_{\max}$
      i. Apply the zoom $Z_j$
      ii. Detect and match SIFT interest points between $I_d^{ref}$ and $I_d^j$
      iii. Estimate the homography $H$ between interest points in $I_d^{ref}$ and $I_d^j$
      iv. Calculate the center $C_d$ of $I_d^{ref}$ in $I_d^j$ by $P_{I_d^j}^k = H \times C_d$
      v. Command the dynamic camera to minimize the distance between $c_d$ and $P_{I_d^j}^k$ until $\left|P_{I_d^j}^k - C_d\right| < \varepsilon$
   d. If the previous step fails, consider $Z_{ref} \leftarrow Z_{ref} + 1$, and go to (1.a)
2. Go to (1) to process $P_{I_s}^{k+1}$ with $Z_{ref} = Z_0$.

### 2.5 Sampling grid

The calibration process involves the matching of interest points extracted with the SIFT algorithm. A regular sampling of the image in the static sensor does not take into

consideration the structure information of the 3D scene. Some of the points of the sampling grid might fall in homogeneous regions, with poor texture information, and cause errors in estimating the homography between the images. To better exploit the 3D scene structure and increase confidence in the learning points, we propose an adaptive sampling strategy which selects more points in textured areas while in homogeneous regions, the mapping will be interpolated from neighboring grid nodes. For a given image $I_s$, SIFT points are detected. Then, a probability density estimated by Parzen windowing (Parzen, 1962) from extracted SIFT interest points is associated to every pixel in $I_s$.(see Fig. 10). The size of the window is taken to be equal to the size of the region of interest used for calibrating at zoom $Z_0$. Two additional constraints are introduced in the sampling method: (i) pixels near the image borders are rejected, (ii) the selected learning nodes must be distributed over the whole image.

Fig. 10. Adaptive grid sampling of the 3-D scene. Left: the source image, middle: image representing the probability density of interest points using the Parzen windowing technique, right: the obtained sampling grid. The size of red circles represents the probability of the node.

### 2.6 Extending the LUT by Thin Plate Spline (TPS) interpolation

The previous learning method enables the determination of a sparse mapping between pixel coordinates in the static camera and angular parameters in the dynamic camera, at a limited number of grid nodes. In order to extend the LUT to all pixels in $I_s$, an approximation is made by using an interpolation function. Thin-Plate-Spline (TPS) interpolation functions, first presented by Bookstein (Bookstein, 1989) are a popular solution to interpolating problems because they give similar results to direct polynomial interpolation, while implicating lower degree polynomials. They also avoid Runge's phenomenon for higher degrees (oscillation between the interpolate points with a large variation).

## 3. Results and discussion

The presented calibration method finds a relation between the coordinates $(x_s, y_s)$ of the point $P_{I_s}$ of $I_s$ and pan-tilt angles such that $P_{I_d}$ coincides with $C_d$. In order to evaluate the accuracy of the method, we seek to estimate the error between the actual position of $P_w$ in $I_d$ and the sought position, e.g. $C_d$. Because the approach is a weak calibration of the camera pair, we have no access to the 3D coordinates of a point in the scene. Consequently,

we used a calibration pattern to estimate the exact coordinates of a point. The pattern is a black ellipse on a white background which is seen in the two cameras (Fig. 11) and easily detectable. The coordinates of the center of gravity of the ellipse was estimated with a subpixellic detector after adaptive thresholding.

Fig. 11. Illustration of the elliptic calibration pattern (surrounded by red) used to evaluate the accuracy of the method of calibration. Left: in the static camera, right: in the dynamic camera.

This calibration method makes a distinction between points learned during the first stage of calibration and interpolated points after the second stage. The learned points serve as a basis for the interpolation function. We present here and evaluation showing, firstly, the accuracy obtained on the learning points, then, taking into account the interpolation stage, the accuracy obtained on any point of the scene observed in the static camera.

### 3.1 Accuracy of learning stage

In order to evaluate the accuracy of the visual servoing process during the calibration, we position the elliptic pattern on a number of nodes on the grid so that its center of gravity coincides with a selected node, and focus the dynamic camera with the learned command parameters. The coordinates of the center of gravity of the ellipse in $I_d$ are determined. Finally, the spatial error between this center of gravity and $C_d$ is estimated and converted to an angular error.

Two dynamic cameras were tested:

- AXIS PTZ 213 network camera, 26x optical zoom, coverage: pan 340°, tilt 100°
- AXIS 233D high-end network dome camera, 35x optical zoom, coverage: pan 360° endless, tilt 180°.

**Results for AXIS 233D dome camera**

The grid points that are considered for evaluation are the points surrounded by black in Fig. 12. The points are sorted in a list, according to their probability density. The initial grid contains 124 nodes. As an example, node 3, 7 and 13 are points where SIFT points density is high. Nodes 43 and 61 present a medium density. The point 93 has a very low density.

One can note (Fig. 13)that the neighborhood of points 3 and 7 represents a region of the scene that exhibits large variations in the 3D geometry in terms of depth (about ten meters). The neighborhood of nodes 13 and 43 presents a greater homogeneity of the 3D geometry of

the scene (a depth of several meters). The neighborhood of points 61 and 93 is mainly composed of a single plane.

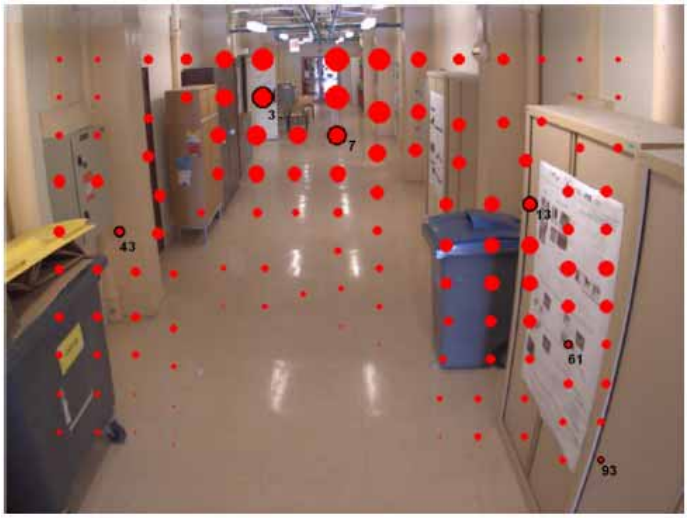


Fig. 12. Learning grid for calibration. The points surrounded by black are used in the discussion on the accuracy of the calibration. The size of the circles represents the probability density of detected SIFT points.


Fig. 13. Neighborhood of six points in $I_s$ selected for evaluation and corresponding to the points surrounded by black in Fig. 12.

The results of the experiments are presented in Table 1.

| | Point 3 | Point 7 | Point 13 | Point 43 | Point 61 | Point 93 |
|---|---|---|---|---|---|---|
| Angular error in pan (degrees) | 0.54° | 0.13° | 0.13° | 0.28° | 0.09° | 0.6° |
| Angular error in tilt (degrees) | 0.02° | 0.3° | 0.16° | 0.29° | 0.08° | 0.32° |

Table 1. Angular errors made in the learning stage of the calibration method.

The method is based on the mapping of interest points that are used to estimate the coordinates of the projected grid point in $I_d$.

The accuracy of the projection estimate depends upon (i) the number of detected and matched SIFT points and (ii) the number of points that verify the homography assumption.
It could be expected that the accuracy of the first points of the grid will be better than the last points because of their higher probability values. The neighborhood of points 13 and 43 and points 61 and 93 are visually similar (Fig.13). The notable difference between the two cases is the SIFT point density in the area of interest. One can notice that the accuracy for points 43 and 93 is lower than for 13 points and 61 although the environment is similar. This result shows the dependence of the calibration accuracy on the 3D scene structure and confirms the interest of this adaptive grid sampling.
However, the accuracy obtained for the points 3 and 7 is much lower than that obtained for items 61, while 3 and 7 own a high probability density. The neighborhood of points 3 and 7 presents sharp disparities in terms of 3D geometry, while the region around point 61 can be better approximated by a plane. This second observation is related to the homography approximation between 3D points of the scene and images in the two sensors.
The corridor is somehow an extreme case because of the large depth variation in the scene, contrary to an office environment for example. Its specific geometry invalidates in some cases the assumption that the learning points locally lie on a plane. Nevertheless, the accuracy achieved with our automatic calibration method remains acceptable in the context of visual surveillance application such as people tracking.

**Results for AXIS 213 PTZ camera**

The deviation due to zoom in this PTZ camera is very important. This means that at a high zoom, a pointed object is no longer entirely visible. For this type of equipment, it is therefore necessary to implement the step size for different zooms. Fig. 14 shows the average error committed at different zooms for a set of points of the learning grid and its associated deviation. One can remark that the errors due to calibration (0.2°-0.3°) are smaller than errors inherent to the zoom mechanism of the PTZ camera (0.6°-0.8°). Our calibration method therefore enables the compensation of the inaccuracy of the camera mechanism.

### 3.2 Accuracy of the interpolation stage

We evaluate here the overall accuracy of the calibration system, including the interpolation step. A number of points distributed over the image and not corresponding to learning points are selected (Fig. 15). The points labelled *a*, *b* and *c* are sampled in the middle of learned points. The points labelled *d* and *e* are chosen in an area with very few learned points because of its homogeneity (ground).
The results of experiments are presented in Table 2.

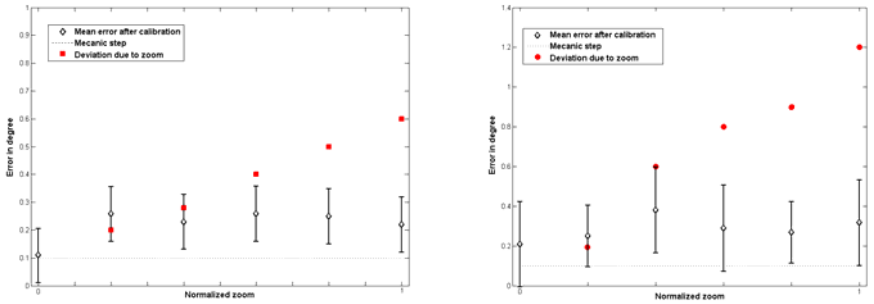


Fig. 14. Results for the accuracy of the method relative to the zoom parameter for the PTZ camera. The errors are represented by their mean and standard deviation, the red dots stand for the observed deviation due to the zoom mechanism. Left: error in degrees of the estimated pan angle, right: error in degrees on the estimated tilt angle.

| | Point *a* | Point *b* | Point *c* | Point *d* | Point *e* |
|---|---|---|---|---|---|
| Angular error in pan (degrees) | 0.12° | 0.26° | 0.13° | 0.07° | 0.5° |
| Angular error in tilt (degrees) | 0.11° | 0.02° | 0.16° | 0.34° | 0.5° |

Table 2. Angular errors for interpolated points.

Fig. 15. Points where the error is measured (blue cross). These points do not belong to the learning grid represented by red dots.

Point *a* shows the best obtained accuracy, since its neighborhood corresponds to a 3D plane. The accuracy of point *b* and *c* is lower. In contrast to point *a*, their neighborhood cannot be easily approximated by a plane. As it was previously shown, the planar constraint affects the obtained accuracy. This influence is also observable at interpolated points. As expected, the accuracy obtained for points taken outside the learning grid (points *d* and *e*) is lower. But it is still acceptable for people tracking applications.

A solution to limit errors due to interpolation could be to increase the number of learning nodes either by using a finer sampling grid or by artificially enriching the scene with textured objects during the off-line calibration process, as shown in Fig. 16.

Fig. 16. Sampling grid in the case of calibration in a corridor. The first row shows the scene observed by the static camera. The second row represents the estimate of the probability density of SIFT points using Parzen windows. The last row shows the obtained sampling grid (the circles size is related to the probability density of the interest point).

## 4. Application to high resolution tracking

An immediate application of our calibration method is to use the pair of cameras as a master-slave system. An object is designated in the static camera and the dynamic camera is commanded to focus on it in order to obtain a higher resolution image. Fig. 17 illustrates the focalization of the dynamic camera using the LUT obtained by the calibration algorithm, for both indoor and outdoor environments. In the outdoor sample, the person and the blue car represent approximately 7x7 pixels in the static camera whereas in the PTZ image, they occupy 300x270 pixels. The object resolution obtained in the PTZ image is suitable for recognition tasks such as gesture recognition, license plate reading or people identification.

Fig. 17. Examples illustrating the direct application of our generic calibration method: indoor (first row) and outdoor (second row) scenes. The images obtained with the dynamic camera (right column) can be used for recognition applications.

We implemented a more elaborate system to automatically detect and track people in the static camera and focus on a particular individual in the dynamic sensor. The detection is carried out by robust and efficient statistical background modelling in the static camera, based on the approach described in (Chen et al., 2007). Detected blobs are then tracked with a Kalman Filter and a simple first order dynamic model, to reinforce spatial coherence of blob/target associations over time. Fig. 18 shows a result of tracking a person along a corridor with our calibrated hybrid dynamic vision system.

## 5. Conclusion

We have proposed a fast and fully automatic learning-based calibration method that determines a complete mapping between the static camera pixels and the command parameters of the dynamic camera, for all values of the zoom. The method encodes in an LUT the following relations:

- $(\alpha_Z, \beta_Z) = \zeta(x_s, y_s)$, for any pixel $(x_s, y_s)$, at a given zoom $Z$,
- $(x_s, y_s) = \zeta^{-1}(\alpha_Z, \beta_Z)$, for any pair $(\alpha_Z, \beta_Z)$, at a given zoom $Z$.

The only requirement is that the observed scene presents sufficient texture information as the methods is based on visual features matching. The obtained results in the corridor

Fig. 18. Tracking a person walking in a corridor with a calibrated hybrid dynamic sensor system.

showed a good accuracy even in the case of high variations of depth in the scene. The knowledge of the complete mapping $\left(\zeta, \zeta^{-1}\right)$ relating the two sensors opens new perspectives for high resolution tracking and pattern recognition in wide areas by collaborative algorithms.

## 6. References


Barreto, J.; Peixoto, P.; Batista, J. & Araujo, H. (1999). Tracking multiple objects in 3D, *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems*, vol. 1, pp. 210–215, ISBN: 0-7803-5184-3, Kyongju, South Korea, October 1999, IEEE Computer Society Washington, DC, USA.

Basu, A. & Ravi, K. (1997). Active camera calibration using pan, tilt and roll. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, Vol. 27, No. 3, (June 1997) pp. 559–566, ISSN: 1083-4419.

Besl, P. J. & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, No. 2, (February 1992) pp. 239–256, ISSN:0162-8828.

Blaszka, T. & Deriche, R. (1995). A Model Based Method for Characterization and Location of Curved Image Features, *Proceedings of the 8th International Conference on Image Analysis and Processing*, pp. 77-82, ISBN: 3-540-60298-4, San Remo, Italy, September 1995, Lecture Notes in Computer Science, Springer.

Bookstein, F. L. (1989). Principal warps: Thin-Plate Splines and the decomposition of deformations. *IEEE Transactions Pattern Analysis Machine Intelligence*, Vol. 11, No. 6, (June 1989) pp. 567–585, ISSN:0162-8828.

Brand, P. & Mohr, R. (1994). Accuracy in image measure, in *Proceedings of the SPIE Conference on Videometrics III*, Vol. 2350, pp. 218- 228, S.F. El-Hakim (Ed.). Boston, Massachusetts, USA.

Chen, Y. & Medioni, G. (1991). Object modelling by registration of multiple range images, *Proceedings of the IEEE International Conference on Robotics and Automation*, Vol. 3, pp. 2724-2729, ISBN: 0-8186-2163-X, Sacramento, CA, USA, April 1991, IEEE Computer Society Washington, DC, USA.

Chen, Y.-T.; Chen, C.-S.; Huang, C.-R. & Hung, Y.-P. (2007). Efficient hierarchical method for background subtraction. *Pattern Recognition*, Vol. 40, No. 10, (October 2007) pp. 2706-2715, ISSN:0031-3203, Elsevier Science Inc., New York, NY, USA.

Collins, R. & Tsin, Y. (1999). Calibration of an outdoor active camera system, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 528–534, ISBN: 0-7695-0149-4 , Fort Collins, CO, USA, June 1999, IEEE Computer Society Washington, DC, USA.

Davis, J. & Chen, X. (2003). Calibrating pan-tilt cameras in wide-area surveillance networks, *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 1, pp. 144–149, ISBN: 0-7695-1950-4, Nice, France, October 2003, IEEE Computer Society Washington, DC, USA.

Fischler, M. A. & Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, Vol. 24, No. 6, (June 1981) pp. 381–395, ISSN: 0001-0782.

Fry, S. N.; Bichsel, M.; Muller, P. & Robert, D. (2000). Tracking of flying insects using pan-tilt cameras. *Journal of Neuroscience Methods*, Vol. 101, No. 1, (August 2000) pp. 59–67, ISSN: 0165-0270.

Gardel, A. (2004). Calibration of a zoom lens camera with pan & tilt movement for robotics. *PhD Thesis*, Université Blaise Pascal, Clermont-Ferrand.

Hartley, R. & Zisserman, A. (2000). *Multiple View Geometry in Computer Vision*, Cambridge University Press, ISBN: 0521623049, New York, NY, USA.

Horaud, R.; Knossow, D. & Michaelis, M. (2006). Camera cooperation for achieving visual attention. *Machine Vision Application*, Vol. 16, No. 6 (February 2006) pp. 1–2, ISSN:0932-8092.

Horaud, R. & Monga, O. (1995). *Vision par ordinateur: outils fondamentaux*, chapter *Géométrie et calibration des caméras*, pp. 139–186, Hermes Science Publications, ISBN: 2-86601-481-2, Paris, France.

Jain, A.; Kopell, D.; Kakligian, K. & Wang, Y.-F. (2006). Using stationary-dynamic camera assemblies for wide-area video surveillance and selective attention, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 537–544, ISBN ~ ISSN:1063-6919 , 0-7695-2597-0, New York, NY, USA, June 2006, IEEE Computer Society Washington, DC, USA.

Lavest, J.-M. & Rives, G. (2003). *Perception visuelle par imagerie vidéo*, chapter *Etalonnage des capteurs de vision*, pp. 23–58, ISBN: 978-2-7462-0662-5, Hermes Science Publications.

Lavest, J.-M.; Viala, M. & Dhome, M. (1998). Do we really need an accurate calibration pattern to achieve a reliable camera calibration*?*, *Proceedings of the 5th European Conference on Computer Vision*, Vol. 1, pp. 158-174, ISBN: 3-540-64569-1, Freiburg, Germany, June 1998, Lecture Notes in Computer Science , Springer.

Lowe, D. G. (1999). Object recognition from local scale-invariant features, *Proceedings of the IEEE International Conference on Computer Vision*, Vol. 2, p. 1150, , September 1999, Kerkyra, Corfu, Greece, ISBN:0-7695-0164-8, IEEE Computer Society Washington, DC, USA.

Mikolajczyk, K. & Schmid, C. (2005). A performance evaluation of local descriptors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 27, No. 10, (June 2003) pp. 1615–1630, ISSN: 1063-6919.

Otsu, N. (1979). A threshold selection method from grey scale histogram. *IEEE Transactions on Systems Man and Cybernetics*, Vol. 9, No. 1, (January 1979) pp. 62-66, ISSN: 0018-9472.

Parzen, E. (1962). On the estimation of a probability density function and mode. *Annals of Mathematical Statistics*, Vol. 33, pp. 1065–1076.

Peuchot, B. (1992). Accurate subpixel detectors, *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Vol. 14, pp.1958-1959, ISBN: 0-7803-0785-2, October 1992, IEEE Computer Society Washington, DC, USA.

Senior, A. W.; Hampapur, A. & Lu, M. (2005). Acquiring multi-scale images by pan-tilt-zoom control and automatic multi-camera calibration, *Proceedings of seventh IEEE Workshops on Application of Computer Vision*, Vol. 1, pp. 433–438, ISBN:0-7695-2271-8-1, Breckenridge, CO, USA, January 2005, IEEE Computer Society Washington, DC, USA.

Stauffer, C. & Tieu, K. (2003). Automated multi-camera planar tracking correspondence modelling, *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 259-266, ISBN: 0-7695-1900-8, Madison, Wisconsin, USA, June 2005, IEEE Computer Society Washington, DC, USA.

Woo, D. & Capson, D. (2000). 3D visual tracking using a network of low-cost pan/tilt cameras, *Proceedings of the Canadian Conference on Electrical and Computer Engineering*, Vol. 2, pp. 884–889, ISBN: 0-7803-5957-7, Halifax, NS, Canada, October 2000, IEEE Computer Society Washington, DC, USA.

Zhou, X.; Collins, R. T.; Kanade, T. & Metes, P. (2003). A master-slave system to acquire biometric imagery of humans at distance, *Proceedings of the ACM international workshop on video surveillance*, pp. 113–120, ISBN:1-58113-780-X, Berkeley, California, USA, ACM New York, NY, USA.