INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Camera cooperation for achieving visual attention

Radu Horaud, David Knossow, and Markus Michaelis

N° 5216

THÈME 3

ISSN 0249-6399 ISRN INRIA/RR--5216--FR+ENG

INRIA
RHÔNE-ALPES

Camera cooperation for achieving visual attention

Radu Horaud, David Knossow, and Markus Michaelis

Thème 3 — Interaction homme-machine,
images, données, connaissances

Projet MOVI

Rapport de recherche n° 5216 — Juin 2004 — 27 pages

Abstract: In this report we address the problem of establishing a computational model for visual attention using cooperation between two cameras. More specifically we wish to maintain a visual event within the field of view of a rotating and zooming camera through the understanding and modelling of the *geometric and kinematic coupling between a static camera and an active camera*. The static camera has a wide field of view thus allowing panoramic surveillance at low resolution. High-resolution details may be captured by a second camera, *provided that it looks in the right direction*. We derive an algebraic formulation for the coupling between the two cameras and we specify the practical conditions yielding a unique solution. We describe a method for separating a foreground event (such as a moving object) from its background while the camera rotates. A set of outdoor experiments shows the two-camera system in operation.

Key-words: video surveillance, visual attention, stereo vision, camera calibration, kinematic calibration, pan-tilt camera head.

This work was partially supported by the ROBEA project PARKNAV. Markus Michaelis was a visiting scientist at INRIA, on leave from Plettac Electronics, Germany

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe, 38330 Montbonnot-St-Martin (France)
Téléphone : +33 4 76 61 52 00 — Télécopie +33 4 76 61 52 52

Coopération entre deux caméras pour l'attention visuelle

Résumé : Dans ce rapport on s'intéresse au problème de l'établissement d'un modèle calculatoire pour décrire l'attention visuelle en utilisant la coopération entre deux caméras. Plus précisément on souhaite maintenir un événement visuel dans le champ de vue d'une caméra, pouvant tourner autour d'elle-même et munie d'un objectif à focale variable, à travers la modélisation du *couplage géométrique et cinématique entre une caméra statique et une caméra active*. La caméra statique possède un champ de vue large permettant ainsi la surveillance panoramique à basse résolution. Les détails perceptibles à haute résolution peuvent être capturés par une seconde caméra *à condition qu'elle regarde dans la bonne direction*. On propose une formulation algébrique pour décrire le couplage entre les deux caméras et on précise les conditions pratiques permettant l'obtention d'une solution unique. On décrit une méthode pour séparer un événement ayant lieu au premier plan (un objet en déplacement, par exemple) de l'arrière plan pendant que la caméra tourne. Un ensemble d'expérimentations faites à l'extérieur illustre le coté opérationnel du système à deux caméras.

Mots-clés : surveillance vidéo, attention visuelle, vision stéréoscopique, calibration de caméra, calibration cinématique, caméra pan-tilt.

1 Introduction

In this paper we address the problem of establishing a computational model for visual attention using cooperation between two cameras. Attention mechanisms may generally be defined as processes that allocate significant computing power to one part or several parts of an image, where information relevant to the task at hand is likely to be found. Therefore, attention processes should encapsulate both top-down and bottom-up visual processes such as (i) the selection of a visual event of interest, (ii) the detection of image features which characterize the selected event, (iii) mechanisms for maintaining these features in the visual field of view, as well as (iv) further analysis such as recognition and interpretation. In this paper we address the problem of maintaining a visual event within the field of view of a camera and the approach that we take consists of monitoring an active camera through the understanding and modelling of the coupling between an active camera and a static camera.

Consider for example the case of a pedestrian or a bicycle rider evolving in an urban environment. They may be viewed as *static objects* in a single image. Nevertheless, in order to take into account the deformable/articulated nature of their shape and motion as well as their time evolution, it is crucial to observe them in videos and therefore consider them as dynamic objects.

Traditional visual attention systems use either an active camera, a binocular active system, or several static cameras. An active camera may rotate, translate, and zoom-in and -out in order to maintain the object of interest within its field of view and in order to compensate for changes in the object's appearance [14], [8], [6], [15]. Binocular devices use controlled camera movements for gaze holding – the two optical axes intersect and produce a zero-disparity surface [3], [2]. Other systems use several static cameras [12]. Static camera configurations have been thoroughly studied from a geometrical point of view [9].

Both single and multiple camera systems have advantages and disadvantages. A single camera is simpler to operate and its motion can be easily controlled with motors. However, it cannot acquire depth information that is useful for scene understanding. Another drawback is that it cannot provide low and high resolution simultaneously. Multiple camera systems have the advantage of being able to acquire potentially richer information provided that the image registration (or correspondence) problem is solved. Active binocular heads try to combine the advantages of controlled motions and multiple camera geometry.

In this paper we propose a solution that combines the advantages of both static and active cameras and of both low- and high-resolution images. One of these cameras is fixed and has a wide field of view, thus allowing surveillance of a wide area both in terms of width and depth of the field of view. Therefore, the image associated with this camera provides a panoramic view while it cannot capture scene details. These scene details are captured by another camera which is equipped with a motor-driven pan and tilt device and with a zoom lens. Therefore, this camera is able to gaze in a specific direction with a specified focal length. At the best of our knowledge the only previous attempt to combine static and active cameras for visual attention and surveillance is described in [16]. The general philosophy

and system architecture described in [16] is very similar to our own approach. We analyse and characterize in detail the fine geometric and mechanical coupling between a static and a rotating camera.

The two-camera video system proceeds as follows. A scene event such as a moving person is first detected and selected using the first (static) camera. Since this camera is static and its field of view covers the whole scene, an event will appear in its associated image sequence as a relatively small object. Well understood and widely developed methods (optical flow, image differentiation, background subtraction, etc.) may be used to *detect* an event occurring in such a region and *track* it over time. However, the resolution associated with this image is not sufficient to properly recognize and interpret the event. The second camera must be controlled in order to dynamically adjust its pan, tilt, and zoom such that the moving object remains in its field of view and such that the object projects onto the image plane at constant size and resolution. Ideally one would like that the camera's degrees of freedom (pan, tilt, and zoom) compensate for changes in appearance due to both viewpoint and depth variations. Once the object of interest has been properly "captured" by the second camera, the latter should be able to track the object using a visual servoing loop which controls the camera's rotations and zoom settings [5].

Such a camera system raises several interesting issues and questions from methodological, computational, and practical points of view. The traditional approach for coupling two or several static cameras based on projective geometry and its associated algebraic and numerical tools is not sufficient. Since one of the cameras is active, both geometrical and mechanical couplings must be considered. Another crucial issue that must be addressed is the stereo correspondence problem. With two static cameras the correspondence problem does not have, in general, a good practical solution because of the inherent ambiguity associated with image-to-image matching. With an active stereo system and under the assumption that a specific object must be selected and tracked, the correspondence problem becomes tractable from a computational point of view. Moreover, stereo correspondence is required only for bootstrapping the attention mechanism. Finally, cooperation between a low-resolution tracker performed with a static camera and a high-resolution tracker performed with an active camera must be properly defined and modelled.

This paper has the following contributions:

- Section 2 *describes and analyses in detail the geometric and kinematic coupling between a static camera and a rotating camera.* The coupling model allows the rotating camera to gaze onto an event selected in the static camera. We derive a mathematical expression for this coupling under the form of a set of polynomial equations and we show that, in the general case, there are several solutions for the pan and tilt angles and these solutions are parameterized by the depth from the static camera to the scene event. We consider the special case where the pan and tilt rotational axes are orthogonal. We specify the practical conditions under which a unique solution exists.

- Section 3 *describes a method for dynamically separating an event (a moving object) from its background.* This is a fundamental step which allows further visual servoing of an active camera. Under the assumptions that the background is static and that the camera is rotating and zooming, there is a 2-D projective mapping between the images in the sequence. We describe a method for robustly estimating this mapping by aligning the grey-levels/colors of image pixels which correspond to the background. This transformation is then used for warping the previous and next frames onto the current frame and for detecting event pixels, i.e., with an apparent image motion that is different than the apparent background motion.
- Section 4 provides an overview of the practical system that is implemented together with some implementation details: camera, stereo, and kinematic calibration, as well as depth estimation with a static-active camera pair. A complete set of experiments is described in detail as well.

2 The coupling between a static and a rotating camera

In this section we consider the geometric and kinematic aspects of the coupling between fixed and rotating cameras. From a geometric point of view, the two cameras act as a stereoscopic device which can be described using the epipolar constraint within a projective geometry framework. From a mechanical point of view, the rotating camera is mounted on a pan and tilt mechanism which has an associated kinematic structure. In order to describe the latter we will adopt a *zero-reference kinematic model.*

Let us denote by $\mathbf{P}_1$ and $\mathbf{P}_2$ the projection matrices associated with the two cameras. A 3-D point M represented by a 4-vector $\boldsymbol{M}$ in projective space is related to its image projections m_1 and m_2 by:

$$\lambda_1 \boldsymbol{m}_1 = \mathbf{P}_1 \boldsymbol{M} \quad (1)$$

$$\lambda_2 \boldsymbol{m}_2 = \mathbf{P}_2 \boldsymbol{M} \quad (2)$$

The non null scalars λ_1 and λ_2 indicate that the projective equality is defined up to a scale factor. They may be interpreted as the *projective depths* along the lines of sight from the centers of projection through the image points m_1 and m_2. For pinhole cameras, the 3×4 projection matrices have the following parameterization:

$$\mathbf{P}_1 = \mathbf{K}_1 \left[\begin{array}{cc} \mathbf{I} & \mathbf{0} \end{array} \right] \quad (3)$$

$$\mathbf{P}_2 = \mathbf{K}_2 \left[\begin{array}{cc} \mathbf{R} & \boldsymbol{t} \end{array} \right] \quad (4)$$

The 3×3 matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ have the intrinsic camera parameters as entries. The rotation $\mathbf{R}$ and the translation $\boldsymbol{t}$ describe the orientation and position of the second camera with respect to the first camera. Without loss of generality we will assume that the first camera is calibrated, therefore matrix $\mathbf{K}_1$ is known. The second camera is calibrated as well up to its focal length f which can vary. The expression of $\mathbf{K}_2$ is:

$$\mathbf{K}_2 = \begin{bmatrix} kf & 0 & u_c \\ 0 & f & v_c \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} k & 0 & u_c \\ 0 & 1 & v_c \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} = \mathbf{K}_2' \mathbf{D}(f, f, 1)$$

With the substitutions $\boldsymbol{m}_1 = \mathbf{K}_1 \boldsymbol{n}_1$ and $\boldsymbol{m}_2 = \mathbf{K}_2' \boldsymbol{n}_2$ we obtain from the equations above:

$$\lambda_2 \boldsymbol{n}_2 = \mathbf{D}(f, f, 1)\,(\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t}) \tag{5}$$

This is the projective epipolar relationship between the camera coordinates $\boldsymbol{n}_1$ and $\boldsymbol{n}_2$ (of m_1 and m_2), the focal length of the active camera f, and the relative position and orientation of the active camera with respect to the static camera, $\boldsymbol{t}$ and $\mathbf{R}$. With $\boldsymbol{n}_2 = (x_2\ y_2\ 1)^\top$ and by denoting $()^i$ the i-th component of a vector, we obtain:

$$\begin{aligned} x_2 &= f \frac{(\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t})_1}{(\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t})_3} \\ y_2 &= f \frac{(\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t})_2}{(\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t})_3} \end{aligned} \tag{6}$$

Without loss of generality we seek a solution which configures the stereo system such that the scene point M is viewed in the center of the image associated with the active camera: $\boldsymbol{n}_2 = (0\ 0\ 1)^\top$. The equations above become:

$$\begin{aligned} (\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t})_1 &= 0 \\ (\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t})_2 &= 0 \end{aligned} \tag{7}$$

Problem formulation. *Given a 3-D point M which is observed in the static camera's image at m_1 with camera coordinates $\boldsymbol{n}_1$ we want to find the position and orientation of the active camera such that M projects onto the active camera's image center.*

In order to solve this problem we must parameterize the rotations and translations of the active camera as a function of the kinematic model associated with it. Therefore we must establish the link between the epipolar geometry and the kinematic model. We will adopt the zero-reference kinematic model for the pan-tilt device. This model allows the user to select a zero-reference or a docking reference for the kinematic chain. We solve for a general pan-tilt kinematic model and we develop a close-form solution for a simplified pan-tilt model. The existence of a unique solution allows to safely apply numerical methods to the general case.

We denote by $\mathbf{T}$ the 4×4 homogeneous matrix:

$$\mathbf{T} = \begin{bmatrix} \mathbf{R} & \boldsymbol{t} \\ \mathbf{0}^\top & 1 \end{bmatrix} \tag{8}$$

We also denote by $\mathbf{T}_0$ the docking or *reference position* of the active camera. From a practical point of view and for stereo calibration purposes, this reference position is chosen such that the two cameras have a common field of view. Let $\mathbf{Q}$ describe the *rigid motion* undergone by the active camera from its reference position to a current position. From Figure 1 one can notice that the following relationship holds:

$$\mathbf{T} = \mathbf{Q}\mathbf{T}_0 \tag{9}$$

2.1 General pan-tilt model

Matrices $\mathbf{T}$ and $\mathbf{T}_0$ have the same structure although the former represents a *rigid motion* while the latter represents a *static* relationship. Matrix $\mathbf{Q}$ represents a *constrained motion*, i.e., a motion undergone by a pan and tilt mechanism. In its most general form this motion can be decomposed as follows (see appendix A):

$$\mathbf{Q} = \mathbf{Q}_2(\alpha, \alpha_0)\mathbf{Q}_1(\beta, \beta_0, \alpha_0) \tag{10}$$

where $\mathbf{Q}_1$ and $\mathbf{Q}_2$ are one-dimensional Lie groups each one describing a rotation around an instantaneous axis: α and β are the *pan* and *tilt* angles parameterizing these motions with α_0 and β_0 being the pan and tilt values associated with the zero-reference position. Each one of these transformations can be written as:

$$\mathbf{Q}_1 = \begin{bmatrix} \mathbf{R}_1 & \boldsymbol{t}_1 \\ \mathbf{0}^\top & 1 \end{bmatrix} = \mathbf{I}_{4\times4} + \sin(\beta - \beta_0)\hat{\mathbf{Q}}_1 + (1 - \cos(\beta - \beta_0))\hat{\mathbf{Q}}_1^2 \tag{11}$$

Matrix $\hat{\mathbf{Q}}_1$ describes the tangent operator associated with the rigid motion; It is composed of a skew-symmetric matrix $\hat{\mathbf{R}}_1$ and a translational velocity vector $\hat{\boldsymbol{t}}_1$ and writes as:

$$\hat{\mathbf{Q}}_1 = \begin{bmatrix} \hat{\mathbf{R}}_1 & \hat{\boldsymbol{t}}_1 \\ \mathbf{0}^\top & 0 \end{bmatrix} \tag{12}$$

It is worthwhile to notice that $\mathbf{Q}_1^{-1}((\beta - \beta_0)) = \mathbf{Q}_1(-(\beta - \beta_0))$ and from equation (11) we obtain that the tangent operator may be estimated from a single motion:

$$\text{trace}\,(\mathbf{Q}_1) = 2\,(1 + \cos(\beta - \beta_0)) \tag{13}$$

and:

$$\hat{\mathbf{Q}}_1 = \frac{1}{2\sin(\beta - \beta_0)}\left(\mathbf{Q}_1 - \mathbf{Q}_1^{-1}\right) \tag{14}$$

By substituting eq. (12) into eq. (11) we obtain:

$$\mathbf{R}_1 = \mathbf{I}_{3\times 3} + \sin(\beta-\beta_0)\hat{\mathbf{R}}_1 + (1-\cos(\beta-\beta_0))\hat{\mathbf{R}}_1^2 \quad (15)$$

$$\boldsymbol{t}_1 = \sin(\beta-\beta_0)\hat{\boldsymbol{t}}_1 + (1-\cos(\beta-\beta_0))\hat{\mathbf{R}}_1\hat{\boldsymbol{t}}_1 \quad (16)$$

There is a similar expression for $\mathbf{Q}_2$. Equation (9) may be written as:

$$\mathbf{R} = \mathbf{R}_2\mathbf{R}_1\mathbf{R}_0 \quad (17)$$

$$\boldsymbol{t} = \mathbf{R}_2\mathbf{R}_1\boldsymbol{t}_0 + \mathbf{R}_2\boldsymbol{t}_1 + \boldsymbol{t}_2 \quad (18)$$

Eq. (7) becomes (the subscripts $()_1$ and $()_2$ denote the first and second vector components):

$$\begin{array}{ll}(\lambda_1\mathbf{R}_2\mathbf{R}_1\mathbf{R}_0\boldsymbol{n}_1 + \mathbf{R}_2\mathbf{R}_1\boldsymbol{t}_0 + \mathbf{R}_2\boldsymbol{t}_1 + \boldsymbol{t}_2)_1 & = 0 \\ (\lambda_1\mathbf{R}_2\mathbf{R}_1\mathbf{R}_0\boldsymbol{n}_1 + \mathbf{R}_2\mathbf{R}_1\boldsymbol{t}_0 + \mathbf{R}_2\boldsymbol{t}_1 + \boldsymbol{t}_2)_2 & = 0\end{array} \quad (19)$$

This a set of of two equations with three unknowns: α, β, and λ_1. We recall that we want to determine the pan and tilt angles such that the event detected at position m_1 in the first image (with camera coordinates $\boldsymbol{n}_1$) appears at position m_2 (with camera coordinates (0 0)) in the second image. The unknown λ_1 is the depth of the observed scene point with respect to the fixed camera. In order to be able to find a solution for the pan and tilt angles we must specify this depth. The practical method for estimating the latter is described in detail in section 4.2. From now on we will assume that λ_1 is known.

In practice it will be more convenient to consider the initial set of three equations, i.e., eq. (5). By substituting equations (17), (18) into this equation and with $\boldsymbol{p} = \lambda_1\mathbf{R}_0\boldsymbol{n}_1 + \boldsymbol{t}_0$ we obtain:

$$\mathbf{R}_1\boldsymbol{p} + \boldsymbol{t}_1 + \mathbf{R}_2^\top\boldsymbol{t}_2 = \mathbf{R}_2^\top\begin{pmatrix}0\\0\\\lambda_2\end{pmatrix} \quad (20)$$

Vector $\boldsymbol{p}$ denotes the coordinates of the observed 3-D point M in the zero-reference camera frame – the "docking" position of the active camera. Therefore we obtain three equations in $\cos(\beta-\beta_0)$, $\sin(\beta-\beta_0)$, $\cos(\alpha-\alpha_0)$, $\sin(\alpha-\alpha_0)$, and $\lambda = \lambda_2$. With the following substitutions we obtain three polynomial equations :

$$\sin(\alpha-\alpha_0) = \frac{2\tan\frac{(\alpha-\alpha_0)}{2}}{1+\tan^2\frac{(\alpha-\alpha_0)}{2}} = \frac{2t_\alpha}{1+t_\alpha^2}$$

$$\cos(\alpha-\alpha_0) = \frac{1-\tan^2\frac{(\alpha-\alpha_0)}{2}}{1+\tan^2\frac{(\alpha-\alpha_0)}{2}} = \frac{1-t_\alpha^2}{1+t_\alpha^2}$$

It is possible to eliminate λ_2 as an unknown between the second and third equations, at the price of increasing the degree of the resulting polynomial. In the general case it will be difficult to analyse the number of admissible solutions of such a set of polynomials [4]. Although in practice these polynomials will be solved using numerical methods, such as the Newton method for finding roots of sets of polynomials, it is crucial to be able to state in advance the exact number of practical solutions.

We denote these sets of solutions by $(\alpha^{(i)}, \beta^{(i)}, \lambda^{(i)})$. They are in the intervals $[\alpha_0 - \pi, \alpha_0 + \pi]$, $[\beta_0 - \pi, \beta_0 + \pi]$ and we must have $\lambda > 0$. So far we considered the most general case. We analyse in detail a simplified pan-tilt device and we show that in this case there is a unique solution. We conclude that the general case also admits a unique solution.

2.2 Simplified pan and tilt model

In the case where the pan and tilt axes are mutually orthogonal the kinematic model of the device is simplified, as described in appendix B. For the purpose of an algebraic analysis and without loss of generality, one may choose $\alpha_0 = \beta_0 = 0$. The matrices become:

$$\mathbf{Q}_1 = \begin{bmatrix} \cos\beta & 0 & \sin\beta & t_1^1 \\ 0 & 1 & 0 & t_2^1 \\ -\sin\beta & 0 & \cos\beta & t_3^1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{21}$$

$$\mathbf{Q}_2 = \begin{bmatrix} 1 & 0 & 0 & t_1^2 \\ 0 & \cos\alpha & -\sin\alpha & t_2^2 \\ 0 & \sin\alpha & \cos\alpha & t_3^2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{22}$$

It follows that eq. (20) becomes:

$$\begin{bmatrix} \cos\beta & 0 & \sin\beta \\ 0 & 1 & 0 \\ -\sin\beta & 0 & \cos\beta \end{bmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} + \begin{pmatrix} t_1^1 \\ t_2^1 \\ t_3^1 \end{pmatrix} + \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\alpha & \sin\alpha \\ 0 & -\sin\alpha & \cos\alpha \end{bmatrix} \begin{pmatrix} t_1^2 \\ t_2^2 \\ t_3^2 \end{pmatrix} = \begin{pmatrix} 0 \\ \lambda_2 \sin\alpha \\ \lambda_2 \cos\alpha \end{pmatrix}$$

which yields the following equations in $\tan\frac{\beta}{2} = t_\beta$, $\tan\frac{\alpha}{2} = t_\alpha$, and $\lambda = \lambda_2$:

$$\begin{aligned} (t_1^1 + t_1^2 - p_1)\, t_\beta^2 + 2p_3\, t_\beta + (t_1^1 + t_1^2 + p_1) &= 0 \\ (t_2^1 - t_2^2 + p_2)\, t_\alpha^2 + 2(t_3^2 - \lambda)\, t_\alpha + (t_2^1 + t_2^2 + p_2) &= 0 \\ (1 + t_\alpha^2)((t_3^1 - p_3)\, t_\beta^2 - 2p_1\, t_\beta + p_3 + t_3^1) &+ \\ (1 + t_\beta^2)((\lambda - t_3^2)\, t_\alpha^2 - 2t_2^2\, t_\alpha - (\lambda - t_3^2)) &= 0 \end{aligned}$$

The first equation has two real solutions for t_β. Indeed, its discriminant is: $\Delta = (p_3)^2 + (p_1)^2 - (t_1^1 + t_1^2)^2$. Obviously the coordinates of vector $\boldsymbol{p}$ have larger values than $t_1^1 + t_1^2$. We recall that vector $\boldsymbol{p}$ represents the coordinates of the observed point M in the zero-reference camera frame. Therefore $\Delta > 0$ and there are two solutions for β in the interval $[-\pi, \pi]$. Only one of these solutions can be achieved in practice, i.e., when the observed point lies in front of the camera. To conclude, the first equation always admits two solutions and only one solution is achievable in practice.

The second equation has two real solutions for t_α as well. Indeed its discriminant is: $\Delta = (t_3^2 - \lambda)^2 - (p_2)^2 + (t_2^2 - t_2^1)^2$. Recall that λ represents the depth from the camera to the observed point and in practical configurations $\lambda >> t_3^2$ and $\lambda >> p_2$. Therefore this equation admits two solutions as well and with the same reasoning as above we conclude that only one solution is achievable in practice.

3 Event/background separation

In the previous section we described the geometric and mechanical coupling allowing the active camera to rotate such that an event detected and tracked with the static camera can be visualized at a higher resolution. In order to be able to analyse this event in more detail, one must properly isolate it from the background.

The event/background separation method is based on two basic assumptions. First, the active camera is supposed to rotate around an axis passing through the center of projection of its associated pinhole representation. Second, it is assumed that the background is static while the event moves with respect to the static scene objects. Therefore, the method described in this section solves for background substraction with a rotating camera.

When a perspective camera undergoes a rotational motion with the instantaneous axis of rotation passing through the center of projection, there is a 2-D projective mapping between the associated images, before and after the motion [10]. This property is straightforward if one applies equations (3) and (4) to the motion of the second camera and with the hypothesis that the translation vector $\boldsymbol{t}$ is null, eq. (5) writes in this case:

$$\boldsymbol{m}^t = \mathbf{K}_2 \mathbf{R}^{t,t-1} \mathbf{K}_2^{-1} \boldsymbol{m}^{t-1} \tag{23}$$

where $\boldsymbol{m}^{t-1}$ and $\boldsymbol{m}^t$ describe the homogeneous coordinates of an image point at times $t-1$ and t, matrix $\mathbf{R}^{t,t-1}$ describes the rotation of the camera from $t-1$ to t, and we denote $\mathbf{H}^{t,t-1} = \mathbf{K}_2 \mathbf{R}^{t,t-1} \mathbf{K}_2^{-1}$.

The rotation undergone in between $t-1$ and t is small enough such that the first order Taylor expansion of eq. (15) holds. This equation writes $\mathbf{R}^{t,t-1} = \mathbf{I}_{3\times 3} + \phi \hat{\mathbf{R}}^{t,t-1}$. Moreover, it is assumed that the focal length is constant between $t-1$ and t. Finally, the homography

can be parameterized as:

$$\mathbf{H}^{t,t-1} = \begin{bmatrix} 1 & -\phi_z^{t,t-1} & f^t\phi_y^{t,t-1} \\ \phi_z^{t,t-1} & 1 & -f^t\phi_x^{t,t-1} \\ -\phi_y^{t,t-1}/f^t & \phi_x^{t,t-1}/f^t & 1 \end{bmatrix} = \begin{bmatrix} 1 & -h_5 & h_1 \\ h_5 & 1 & h_2 \\ h_3 & h_4 & 1 \end{bmatrix} \tag{24}$$

It is worthwhile to notice that the scalars ϕ_x, ϕ_y, and ϕ_z are the projections of $\phi\boldsymbol{k}$ onto the three coordinate axes of the camera frame, where $\boldsymbol{k}$ is the unit direction vector of the instantaneous axis of rotation and ϕ is the instantaneous angle of rotation. The latter is to be accurately estimated in order to align the images. It is therefore different than the pan and tilt angles which are outputs of the camera control device.

The relationship between $\boldsymbol{m}^{t-1}$ and $\boldsymbol{m}^t$ above is valid for static scene points. In the past this was used in combination with an outlier rejection technique in order to segment the image into two layers: a static layer corresponding to a static background and a dynamic layer corresponding to moving objects – a foreground. However, such a strategy is generally based on robust statistical methods applied to a single rotating camera.

With the camera configuration being used here the segmentation algorithm is greatly simplified. Indeed, moving objects are detected as events in the image associated with the static camera. The camera coupling allows to predict the main event under investigation, to place the second camera, and to adjust its settings, such that this event is centered with respect to the active camera coordinate frame. Therefore, a major advantage associated with this two-camera configuration is that a robust statistical method *is not required*. This is best shown on Figure 2.

The separation between an event and its background is therefore based on (i) aligning the images based on the static background and (ii) on comparing them, pixel by pixel. The event detection performed with the low-resolution static image bootstraps this process.

From now on we consider the images associated with the active camera and we assume that these images are segmented into two regions: foreground $\mathcal{F}$ and background $\mathcal{B}$. In order to find the homography which aligns the backgrounds between times t and $t-1$, the following error must be minimized:

$$E_{\min} = \min_{h_i} \sum_{\boldsymbol{m}\in\mathcal{B}} \|\mathcal{I}^{t-1}(\Psi(\boldsymbol{m}^{t-1})) - \mathcal{I}^t(\Psi(\mathbf{H}^{t,t-1}\boldsymbol{m}^{t-1}))\|^2 \tag{25}$$

The function $\Psi()$ denotes the non linear mapping from homogeneous to Euclidean coordinates, $\Psi(m_1, m_2, m_3) = (m_1/m_3,\ m_2/m_3)^\top$. Various methods were developed in the past for solving this non-linear minimization problem [11], [19], [1]. Notice that with the special homography introduced by eq. (24) the number of parameters to be estimated is equal to 5. Moreover, the use of a controlled camera allows proper initialization of the rotational angle values.

Once such a homography is estimated, it optimally aligns the backgrounds. The statistics associated with the actual minimization result ($E_{\min}$) allows one to associate a probability

of background with each pixel. These statistics can be improved if a background image is incrementally built as is done in [1]. Eventually one may use a decision rule in order to decide whether a pixel belongs to the background or to the foreground [7]. In practice such an approach will not perform as well as expected simply because background and foreground image regions may have similar grey-level or color values.

Therefore, to further refine the foreground area we proceed by pixel-to-pixel comparison between three images at times $t-2$, $t-1$, and t. The difference between two pixels corresponding to two aligned images writes:

$$\mathcal{D}^{t,t-1}(\Psi(\boldsymbol{m}^{t-1})) = \left(\mathcal{I}^{t-1}(\Psi(\boldsymbol{m}^{t-1})) - \mathcal{I}^{t}(\Psi(\mathbf{H}^{t,t-1}\boldsymbol{m}^{t-1}))\right)^2 \tag{26}$$

There is a similar expression for $\mathcal{D}^{t-1,t-2}(\Psi(\boldsymbol{m}^{t-2}))$ where the mapping $\boldsymbol{m}^{t-1} = \mathbf{H}^{t-1,t-2}\boldsymbol{m}^{t-2}$ holds for the background. As already mentioned, statistics associated with the minimization of eq.(25) allows the estimation of a threshold s such that the following simple decision rule is used: A pixel $\boldsymbol{m}^t$ belongs to the foreground if:

$$\mathcal{D}^{t,t-1}(\Psi(\boldsymbol{m}^{t-1})) \geq s \text{ and } \mathcal{D}^{t-1,t-2}(\Psi(\boldsymbol{m}^{t-2})) \geq s$$

4 Methodology, implementation, and experiments

High-quality pan-tilt cameras available today can achieve a precision of about 0.05^0 in pan and tilt. The precision to be reached in practice, using a calibrated camera setup such as the one described in this section, is of the order of 0.1^0. Consider for example a field of view with an aperture angle of about 2^0. At 100 meters the width of the field of view is 3.5 meters and therefore the precision is of the order of 0.2 meters. This is sufficient to gaze and zoom onto a person across a football stadium, or onto a bicycle, a pedestrian, and/or a car in a typical traffic scenario. This overall precision – 0.1^0 – can be achieved only if the system is properly calibrated.

Another important ingredient of such a two-camera device is the control of the active camera such that it continuously looks towards the object of interest and maintains its gaze such that this object appears at its image center. This gaze-control process requires two steps: initialization and visual servoing.

The two-camera visual attention system, once calibrated, proceeds as follows:

- Initialization:
  - A scene object is detected and tracked over time with the wide-lens static camera;
  - The active camera is controlled such that this scene object falls within its field of view;

  - The two-camera system estimates the depth to the scene object;
  - The active camera's pan and tilt mechanisms is controlled to gaze onto the scene object;
  - As the active camera moves, three consecutive images are used to separate the moving object (foreground pixels) from the static background. The center of gravity of the foreground pixels is estimated.

- Visual servoing loop:
  - As the scene object moves, the active camera is controlled such that the center of gravity of the foreground pixels is maintained at the image center;
  - Three consecutive images are used to separate the moving object from the background and to update its center of gravity (as the object changes its size, shape, etc.).

4.1 Calibration of the two-camera device

The camera cooperation method described in this paper effectively works in practice only on the premises that the geometric and kinematic parameters of the two camera device are properly estimated. This is performed by the following steps:

1. *Intrinsic camera calibration.* The intrinsic parameters of both cameras, i.e., $\mathbf{K}_1$ and $\mathbf{K}_2$ in eqs. (3), (4), are calibrated using a classical camera calibration process as described in detail in [17].

2. *Stereo calibration.* When the active camera is in its docking or zero-reference position, the two cameras may be viewed as a standard stereoscopic pair characterized either by the rotation and translation between the two camera frames (stereo calibration) or by the epipolar geometry (weak stereo calibration). The method described in [18] allows for an accurate stereo calibration by moving a 3-D pattern in front of the cameras. Eventually, the matrix $\mathbf{R}_0$ and the vector $\boldsymbol{t}_0$ characterizing the camera setup in its docking position are evaluated.

3. *Kinematic calibration.* The active camera is mounted onto a pan and tilt device – two coupled rotational motions. Kinematic calibration consists in estimating the tangent operators associated with these constrained motions, i.e., $\hat{\mathbf{Q}}_1$ in eq. (11). The pan-tilt kinematic model is formally described in appendix A. The kinematic calibration procedure is described in detail in appendix C.

4.2 Depth estimation

The method described in sections 2.1 and 2.2 returns a unique set of values for the pan and tilt angles provided that an estimation of the depth from the static camera to a scene object is available – λ_1. In this section we describe a practical technique for estimating the depth to a scene object. This involves the following steps:

1. Detect this object in the static image and locate its center, say m_1;
2. Control the active camera such that it looks in the right direction and therefore the epipolar line associated with m_1 is visible in its image, and
3. Search along this epipolar line in order to find the best match of m_1, say m_2, and estimate the depth to the scene object.

Let us suppose that this object is detected and located in the fixed image and let m_1 with camera coordinates $\boldsymbol{n}_1$ be the image of its center. The scene object lies somewhere along the line of sight associated with this image point, i.e., Figure 3.

Let $\lambda_{\min}$ and $\lambda_{\max}$ be the minimum and maximum expected depth values along this line of sight such that $\lambda_{\min} \leq \lambda_1 \leq \lambda_{\max}$. We associate two points with these depth values, $M_{\min}$ and $M_{\max}$. They project onto the active camera's image plane at $m_{\min}$ and $m_{\max}$. These image-plane points lie on the epipolar line associated with m_1. We seek a position, an orientation, *and a focal length* for the active camera such that the epipolar line-segment between $m_{\min}$ and $m_{\max}$ is actually visible in the image.

We constrain this epipolar line-segment to be a horizontal image line passing through the image center, i.e., the coordinates of $m_{\min}$ and $m_{\max}$ verify: $\boldsymbol{n}_{\min} = (c, 0, 1)^\top$ and $\boldsymbol{n}_{\max} = (-c, 0, 1)^\top$, where $2c$ corresponds to the image width. The image coordinates of these points verify eq. (6):

$$c = f\frac{(\lambda_{\min}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_1}{(\lambda_{\min}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_3} \qquad 0 = f\frac{(\lambda_{\min}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_2}{(\lambda_{\min}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_3}$$

$$-c = f\frac{(\lambda_{\max}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_1}{(\lambda_{\max}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_3} \qquad 0 = f\frac{(\lambda_{\max}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_2}{(\lambda_{\max}\mathbf{R}\boldsymbol{n}_1+\boldsymbol{t})_3}$$

In order to solve these equations and estimate $\mathbf{R}$, $\boldsymbol{t}$, and f, we recall that the rotation matrix and the translation vector can be parameterized by the pan and tilt angles α, β and by the stereo calibration parameters $\mathbf{R}_0$ and $\boldsymbol{t}_0$, e.g., eqs (17) and (18). Nevertheless, this parameterization does not allow proper alignment because the active camera cannot rotate around its optical axis. For this reason we introduce a third rotation allowing a *virtual* rotation of the active camera around its z-axis, $\mathbf{R}_3(\gamma)$.

Therefore, there are four equations in four unknowns, f, α, β, and γ. A solution can be found using the Newton's method for solving a set of polynomials. Notice that for each

point-to-point correspondence and for a given depth value, there is a unique solution in α and β. Hence, one can use the known triplets $\boldsymbol{n}_1, \boldsymbol{n}_{\min}, \lambda_{\min}$ and $\boldsymbol{n}_1, \boldsymbol{n}_{\max}, \lambda_{\max}$ to find initial values for the pan and tilt angles and guarantee that the active camera gazes in the right direction.

The active camera is controlled to zoom and rotate in order to reach the solution found above up to a rotation γ around its optical axis. Eventually, standard stereo techniques are applied in order to find the best match along the epipolar line and to estimate the depth to the scene object.

4.3 Experiments

Figures 4, 5, and 6 summarize a full set of experiments. The stereo baseline between the static and active cameras is of the order of 1 meter. The cameras observe an outdoor environment. The frames which are shown correspond to 6 samples out of a 550-frame image sequence.

The object of interest (a walking person) is first detected in the image associated with the static camera. Given minimum and maximum depth estimates, from the static camera to that person, the active camera rotates and zooms such that the person falls within its field of view. Since the camera couple is calibrated, it is possible to predict an epipolar line, to search for a match along this line, and to estimate the actual depth from the event to the static camera.

The pan and tilt values allowing to place the person's center of gravity at the image center are estimated and the active camera's mechanism is controlled to actually place the person in its center. A region of interest is defined around the moving object.

Finally, a visual servoing loop controls the camera in order to maintain the object at the image center. While the camera rotates, the moving foreground is separated from the static background, e.g., figure 6.

5 Conclusion

In this paper we addressed the problem of coupling two cameras in order to achieve visual attention. The first camera is static and it has a wide field of view. Therefore it is able to capture, at low resolution, such events as moving articulated/deformable objects. The second camera is mounted onto a rotating device with two degrees of freedom. Moreover it has a narrow field of view – of the order of 2 degrees. Therefore it is able to provide a high-resolution image of a scene object, provided that the latter falls within its field of view.

We analyzed in detail the geometric and kinematic coupling between a static and a rotating camera. We derived a solution for this coupling both for a general kinematic

mechanism and for a simple pan-tilt model. We showed that under practical conditions there is a unique solution allowing to rotate the camera such that it gazes towards an object scene. This solution is parameterized by a depth parameter (the distance from the static camera to the object) and we described a practical solution to estimate this depth.

Once the object of interest lies along the active camera's optical center a visual servoing loop can be used to control the camera's rotational degrees of freedom. Moreover, knowing the approximate location of the object in the image we showed that is possible to segment it from the background.

The vast majority of visual surveillance and visual attention systems use a single camera. Cooperation between static and active cameras is an essential step forward allowing to rapidly analyse an event at low resolution and to switch to high resolution if further recognition and interpretation steps are necessary.

A The pan-tilt kinematic model

In this appendix we formally define the rotational mechanism associated with the active camera. First we consider the most general kinematic model. We adopt the zero-reference kinematic representation. The angle associated with the "tilt" rotation is denoted by β. The angle associated with the "pan" rotation is denoted by α. The kinematic chain is composed of five Euclidean frames and four rigid transformations between these frames, see Figure 7:

- Frame #5 is attached to a fixture, it is equivalent to the "base" of the device;
- Frame #4 is a moving frame rotating around frame #5; This tilt rotation is denoted by $\mathbf{T}_1$ which is a 4×4 homogeneous matrix denoting a rigid transformation;
- Frame #3 is rigidly attached to frame #4 through the fixed transformation $\mathbf{L}_1$;
- Frame #2 is a moving frame rotating around frame #3; This pan rotation is denoted by $\mathbf{T}_2$;
- Frame #1, or the camera frame, is rigidly attached to frame #2 through the fixed transformation $\mathbf{L}_2$

The coordinates of the physical point M (observed by the camera) can be written in camera coordinates, $\boldsymbol{M}^{(1)}$, as well as in fixture coordinates, $\boldsymbol{M}^{(5)}$. Obviously we have:

$$\boldsymbol{M}^{(1)}(\alpha,\beta) = \mathbf{L}_2\mathbf{T}_2(\alpha)\mathbf{L}_1\mathbf{T}_1(\beta)\boldsymbol{M}^{(5)} \tag{27}$$

The same formula holds for a docking position which is referred to as the zero-reference and which is characterized by *fixed* values for the two angles:

$$\boldsymbol{M}^{(1)}(\alpha_0,\beta_0) = \mathbf{L}_2\mathbf{T}_2(\alpha_0)\mathbf{L}_1\mathbf{T}_1(\beta_0)\boldsymbol{M}^{(5)} \tag{28}$$

By eliminating $\boldsymbol{M}^{(5)}$ between these equations we obtain the *zero-reference* kinematic model of the active camera, i.e., eq. (10):

$$\boldsymbol{M}^{(1)}(\alpha,\beta) = \mathbf{Q}_2(\alpha,\alpha_0)\mathbf{Q}_1(\beta,\beta_0,\alpha_0)\boldsymbol{M}^{(1)}(\alpha_0,\beta_0) \tag{29}$$

The reference frames have been appropriately defined such that (without loss of generality) the transformations $\mathbf{T}_1$ and $\mathbf{T}_2$ can be written in a canonical form, i.e., rotation around the local z-axis:

$$\mathbf{T}_1(\beta) = \begin{bmatrix} \cos\beta & -\sin\beta & 0 & 0 \\ \sin\beta & \cos\beta & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{30}$$

These matrices form a one-dimensional Lie group with $\mathbf{T}_1^{-1}(\beta) = \mathbf{T}_1(-\beta)$. Therefore, from the equations above we obtain the following expressions for $\mathbf{Q}_2$ and $\mathbf{Q}_1$:

$$\mathbf{Q}_2(\alpha,\alpha_0) = \mathbf{L}_2\mathbf{T}_2(\alpha-\alpha_0)\mathbf{L}_2^{-1} \tag{31}$$

$$\mathbf{Q}_1(\beta,\beta_0,\alpha_0) = \underbrace{\mathbf{L}_2\mathbf{T}_2(\alpha_0)\mathbf{L}_1}_{\mathbf{U}_1}\,\mathbf{T}_1(\beta-\beta_0)\,\underbrace{\mathbf{L}_1^{-1}\mathbf{T}_2^{-1}(\alpha_0)\mathbf{L}_2^{-1}}_{\mathbf{U}_1^{-1}} \tag{32}$$

Since matrices $\mathbf{Q}_i$ and $\mathbf{T}_i$ are related by similarity transformations, it follows that both $\mathbf{Q}_1$ and $\mathbf{Q}_2$ form one-dimensional Lie groups as well. It is well known, [13], that these groups can be parameterized using their Lie algebra and their angle of rotation, i.e., eq. (11).

B Simple pan-tilt model

In the case of a simplified model it is assumed that the pan and tilt axes are mutually perpendicular. In this case, matrices $\mathbf{L}_1$ and $\mathbf{L}_2$ are pure translational offsets:

$$\mathbf{L}_1 = \begin{bmatrix} 0 & 0 & 1 & l_1^1 \\ 1 & 0 & 0 & l_2^1 \\ 0 & 1 & 0 & l_3^1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{33}$$

$$\mathbf{L}_2 = \begin{bmatrix} 0 & 0 & 1 & l_1^2 \\ 1 & 0 & 0 & l_2^2 \\ 0 & 1 & 0 & l_3^2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{34}$$

We obtain:

$$\mathbf{Q}_2 = \begin{bmatrix} 1 & 0 & 0 & t_1^2 \\ 0 & \cos(\alpha-\alpha_0) & -\sin(\alpha-\alpha_0) & t_2^2 \\ 0 & \sin(\alpha-\alpha_0) & \cos(\alpha-\alpha_0) & t_3^2 \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{35}$$

$$\mathbf{Q}_1 = \mathbf{U}_1 \begin{bmatrix} \cos(\beta-\beta_0) & -\sin(\beta-\beta_0) & 0 & 0 \\ \sin(\beta-\beta_0) & \cos(\beta-\beta_0) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \mathbf{U}_1^{-1} \tag{36}$$

with:

$$\mathbf{U}_1 = \begin{bmatrix} 0 & 1 & 0 & l_3^1 + l_1^2 \\ -\sin\alpha_0 & 0 & \cos\alpha_0 & l_1^1\cos\alpha_0 - l_2^1\sin\alpha_0 + l_2^2 \\ \cos\alpha_0 & 0 & \sin\alpha_0 & l_1^1\sin\alpha_0 + l_2^1\cos\alpha_0 + l_3^2 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

C Kinematic calibration

Kinematic calibration consists in estimating the Lie algebras associated with the matrices $\mathbf{Q}_1$ and $\mathbf{Q}_2$ formally defined in appendix A. Each one of these matrices form a one-parameter Lie group such that $\mathbf{Q}_1(\beta_1)\mathbf{Q}_1(\beta_2) = \mathbf{Q}_1(\beta_1+\beta_2)$. Moreover, once a reference frame is being chosen, the tangent operator (or the Lie algebra) remains fixed. Therefore, the kinematic calibration process consists in finding a numerical estimate of $\hat{\mathbf{Q}}_1$ and of $\hat{\mathbf{Q}}_2$, i.e., eq. (14). For that purpose we consider again eq. (29). Notice that the transformation from position α_1, β_1 to position α_2, β_2 writes:

$$\mathbf{Q}_{\alpha_1\to\alpha_2,\beta_1\to\beta_2} = \mathbf{Q}_2(\alpha_2)\mathbf{Q}_1(\beta_2-\beta_1)\mathbf{Q}_2(\alpha_1)$$

Let the pan-tilt device perform two one-parameter motions: a motion from α_1 to α_2 and another motion from β_1 to β_2. From the equation above we obtain:

$$\mathbf{Q}_2(\alpha_2-\alpha_1) = \mathbf{Q}_{\alpha_1\to\alpha_2,\beta_1} \tag{37}$$

$$\mathbf{Q}_1(\beta_2-\beta_1) = \mathbf{Q}_2(-\alpha_1)\mathbf{Q}_{\alpha_1,\beta_1\to\beta_2}\mathbf{Q}_2(\alpha_1) \tag{38}$$

In practice the kinematic calibration proceeds as follows:

Step 1: Move the device in the α_1, β_1 position;

Step 2: Using camera calibration tools, estimate the external camera parameters, i.e., the position and orientation of the camera frame with respect to a calibration fixture expressed as a rigid transformation $\mathbf{T}(\alpha_1, \beta_1)$;

Step 3: Move the device in the α_2, β_1 position;

Step 4: Repeat Step 2 for this position and estimate $\mathbf{T}(\alpha_2, \beta_1)$;

Step 5: Move the device in the α_1, β_2 position;

Step 6: Repeat Step 2 for this position and estimate $\mathbf{T}(\alpha_1, \beta_2)$;

Step 7: Compute $\mathbf{Q}_{\alpha_1 \to \alpha_2, \beta_1} = \mathbf{T}(\alpha_2, \beta_1)\mathbf{T}(\alpha_1, \beta_1)^{-1}$;

Step 8: Compute $\hat{\mathbf{Q}}_2$ from $\mathbf{Q}_2(\alpha_2 - \alpha_1)$ using eq. (14);

Step 9: Compute $\mathbf{Q}_{\alpha_1, \beta_1 \to \beta_2} = \mathbf{T}(\alpha_1, \beta_2)\mathbf{T}(\alpha_1, \beta_1)^{-1}$;

Step 10: Compute $\mathbf{Q}_2(\alpha_1)$, $\mathbf{Q}_2(-\alpha_1)$, and $\mathbf{Q}_1(\beta_2 - \beta_1)$, and

Step 11: Compute $\hat{\mathbf{Q}}_1$ from $\mathbf{Q}_2(\beta_2 - \beta_1)$ using eq. (14);

References

[1] A. Bartoli, N. Dalal, and R. Horaud. Motion panoramas. Technical Report RR-4771, INRIA, February 2003.

[2] J. Batista, P. Peixoto, and H. Araujo. Real-time active visual surveillance by integrating peripheral motion detection with foveated tracking. In *IEEE Workshop on Visual Surveillance*, Mumbai, India, 1998.

[3] D. Coombs and C. Brown. Real-time binocular smooth pursuit. *International Journal of Computer Vision*, 11(2):147–164, October 1993.

[4] D. Cox, J. Little, and D. O'Shea. *Using Algebraic Geometry*. Springer, 1998.

[5] A. Crétual and F. Chaumette. Application of motion-based visual servoing to target tracking. *Int. Journal of Robotics Research*, 20(11):878–890, November 2001.

[6] K. Daniilidis, C. Krauss, M. Hansen, and G. Sommer. Real time tracking of moving objects with an active camera. *Real Time Imaging*, 4(1):3–20, February 1998.

[7] F. Dufaux and F. Moscheni. Background mosaicking for low bit rate video coding. In *Proceedings IEEE International Conference on Image Processing*, volume 1, pages 673–676, Lausanne, Switzerland, September 1996.

[8] J. A. Fayman, O. Sudarsky, E. Rivlin, and Rudzsky. M. Zoom tracking and its applications. *Machine Vision and Applications*, 13(1):25–37, August 2001.

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

[10] R. I. Hartley. Self-calibration from multiple views with a rotating camera. In *Proc. Third European Conference on Computer Vision*, pages 471–478, Stockholm, Sweden, May 1994.

[11] M. Irani and P. Anandan. About direct methods. In B. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, number 1883 in LNCS, pages 267–277, Corfu, Greece, July 1999. Springer-Verlag.

[12] F Martin and R. Horaud. Multiple camera tracking of rigid objects. *International Journal of Robotics Research*, 21(2):97–113, February 2002.

[13] J. M. McCarthy. *Introduction to Theoretical Kinematics*. MIT Press, Cambridge, 1990.

[14] D. Murray and A. Basu. Motion tracking with an active camera. *IEEE Transactions on Pattern Analysis and Machine Intteligence*, 16(5):449–459, May 1994.

[15] D.W. Murray, K.J. Bradshaw, P.F. McLauchlan, I.D. Reid, and P.M. Sharkey. Driving saccade to pursuit using image motion. *International Journal of Computer Vision*, 16(3):205–228, November 1995.

[16] P. Peixoto, J. Batista, and H. Araujo. Integration of information from several vision systems for a common task of surveillance. In *6th Int. Workshop on Intelligent Robotics Systems*, Edinburgh, UK, 1998.

[17] Matthieu Personnaz and Radu Horaud. Camera calibration: estimation, validation and software. Technical Report RT-0258, INRIA Rhone Alpes, Grenoble, March 2002.

[18] Matthieu Personnaz and Peter Sturm. Calibration of a stereo-vision system by the non-linear optimization of the motion of a calibration object. Technical Report RT-0269, INRIA, September 2002.

[19] H.-Y. Shum and R. Szeliski. Systems and experiment paper: Construction of panoramic mosaics with global and local alignment. *International Journal of Computer Vision*, 36(2):101–130, February 2000.

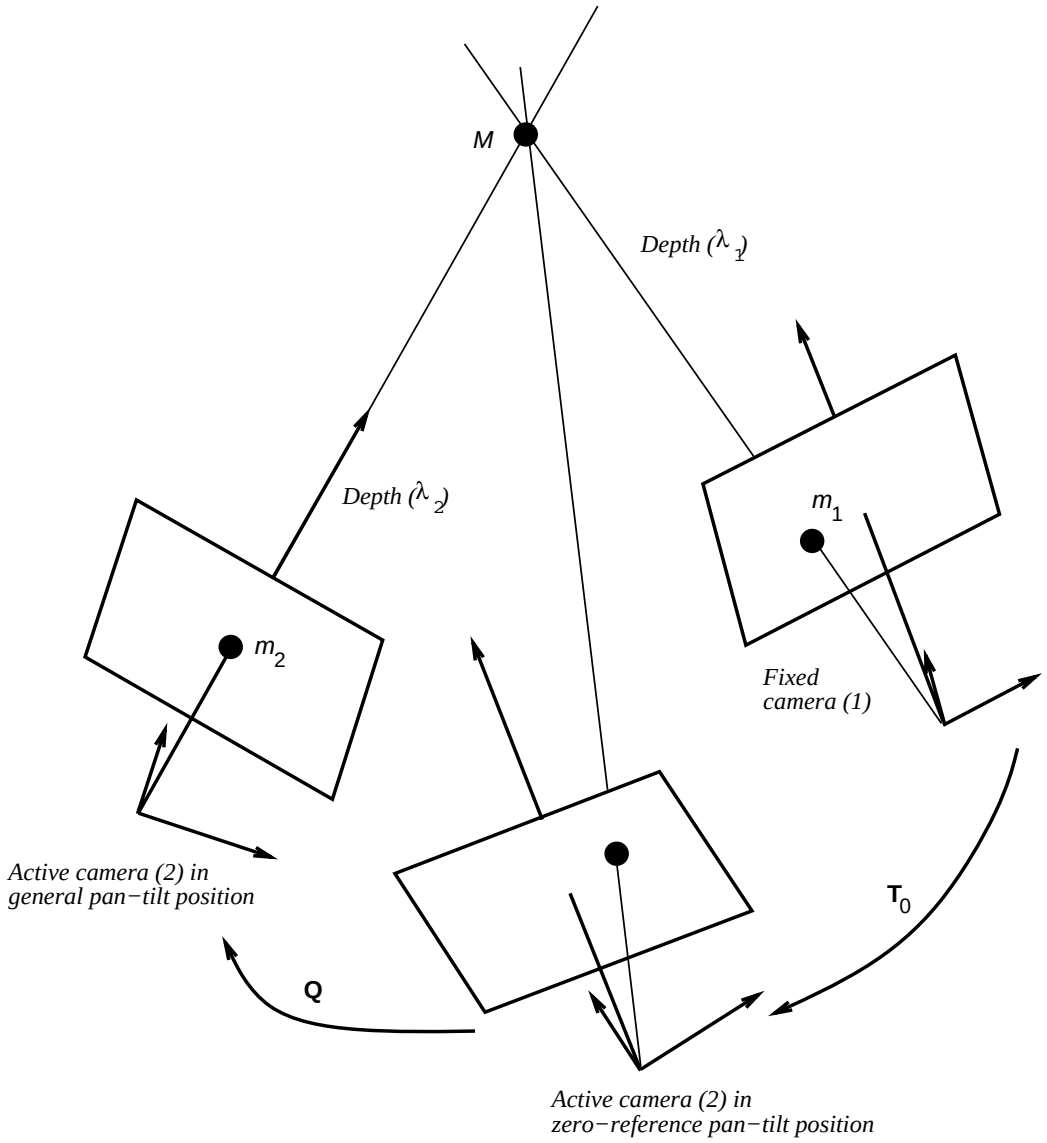

Figure 1: The active camera has a docking or a zero-reference position. Both the stereo (matrix $\mathbf{T}_0$) and kinematic (matrix $\mathbf{Q}$) calibrations are performed with respect to this reference position.

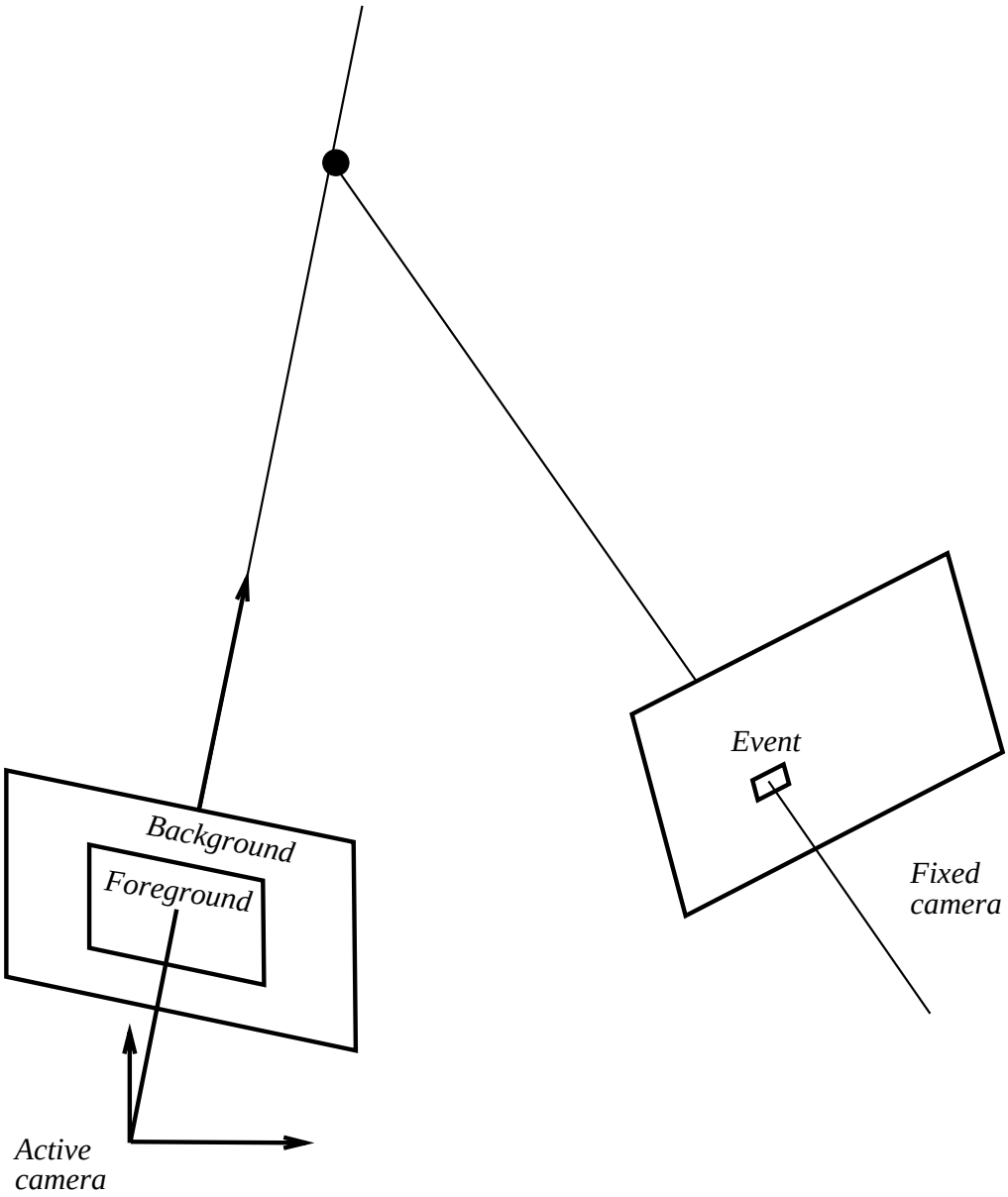

Figure 2: The coupling between the cameras allows one to associate *foreground* and *background* regions with the active camera's image. The event, which is predicted in the static camera at low resolution, must lie in the foreground region associated with the active camera's image sequence.

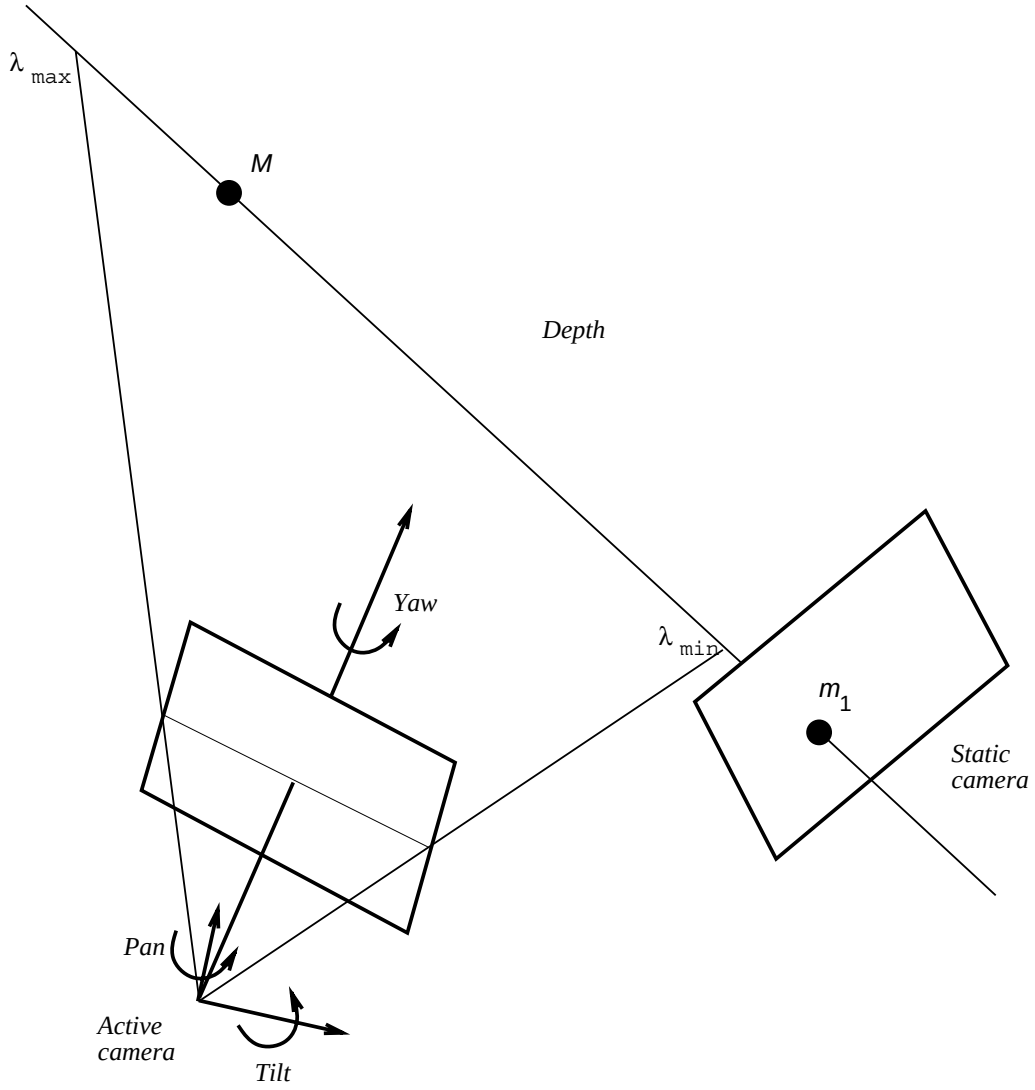

Figure 3: In order to estimate the depth to the point M, the active camera must see this point. The degrees of freedom of the active camera – pan, tilt, yaw, and focal length – are estimated such that the line of sight associated with image point m_1 is seen in the active image plane.

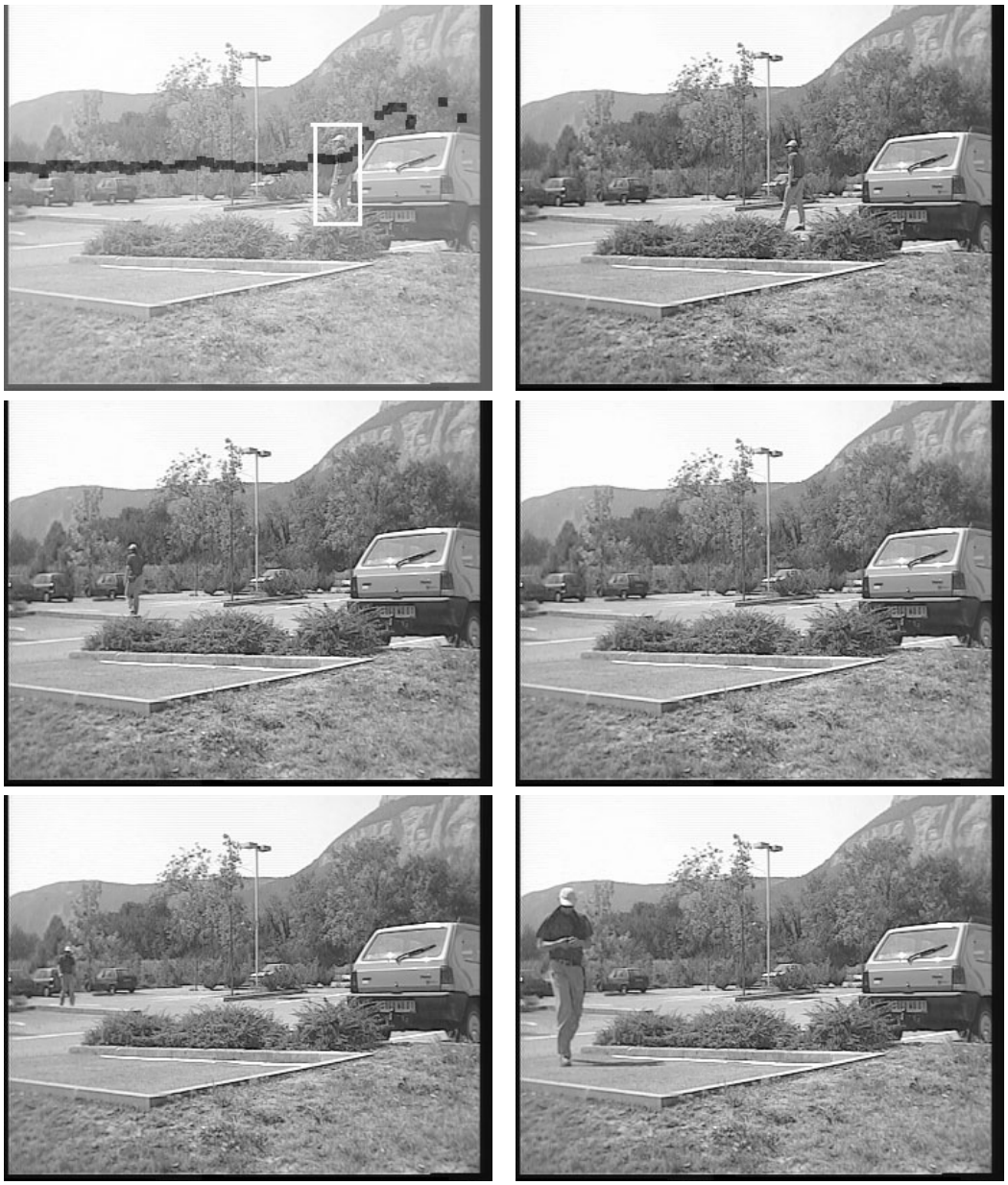

Figure 4: This figures shows 6 frames out of a 550-frame image sequence available with the static camera, namely frames 97, 144, 184, 284, 384, and 527. The first image shows the trajectory of the moving person as well as a bounding box around this person at some time instant. Notice that the person temporarily disappears from the field of view of this camera.

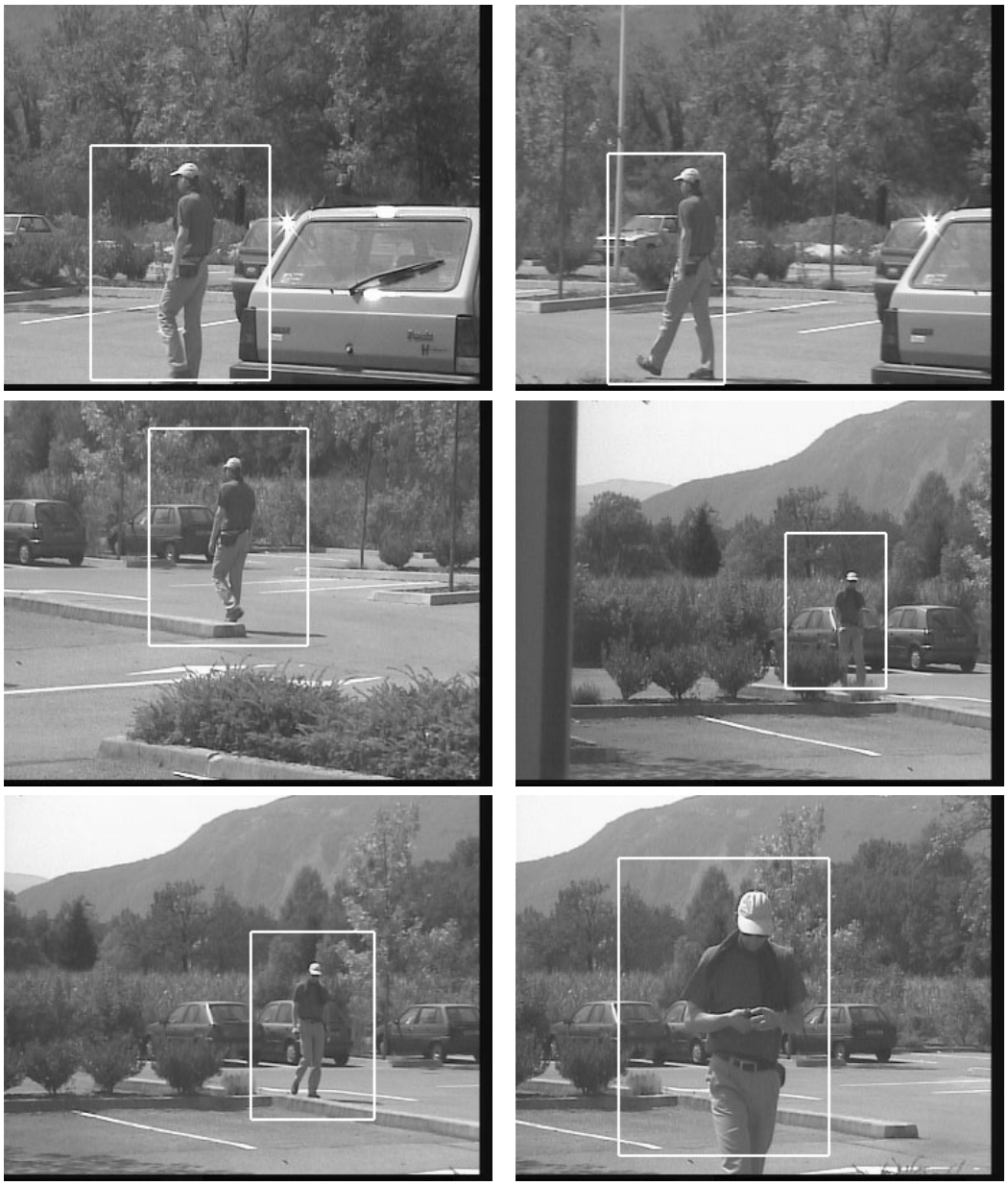

Figure 5: This figure shows the output of the active camera. The object of interest (the person) has its center of gravity approximatively aligned with the image center. The visual servoing loop allows to keep track of the object even if it disappears from the field of view of the static camera.

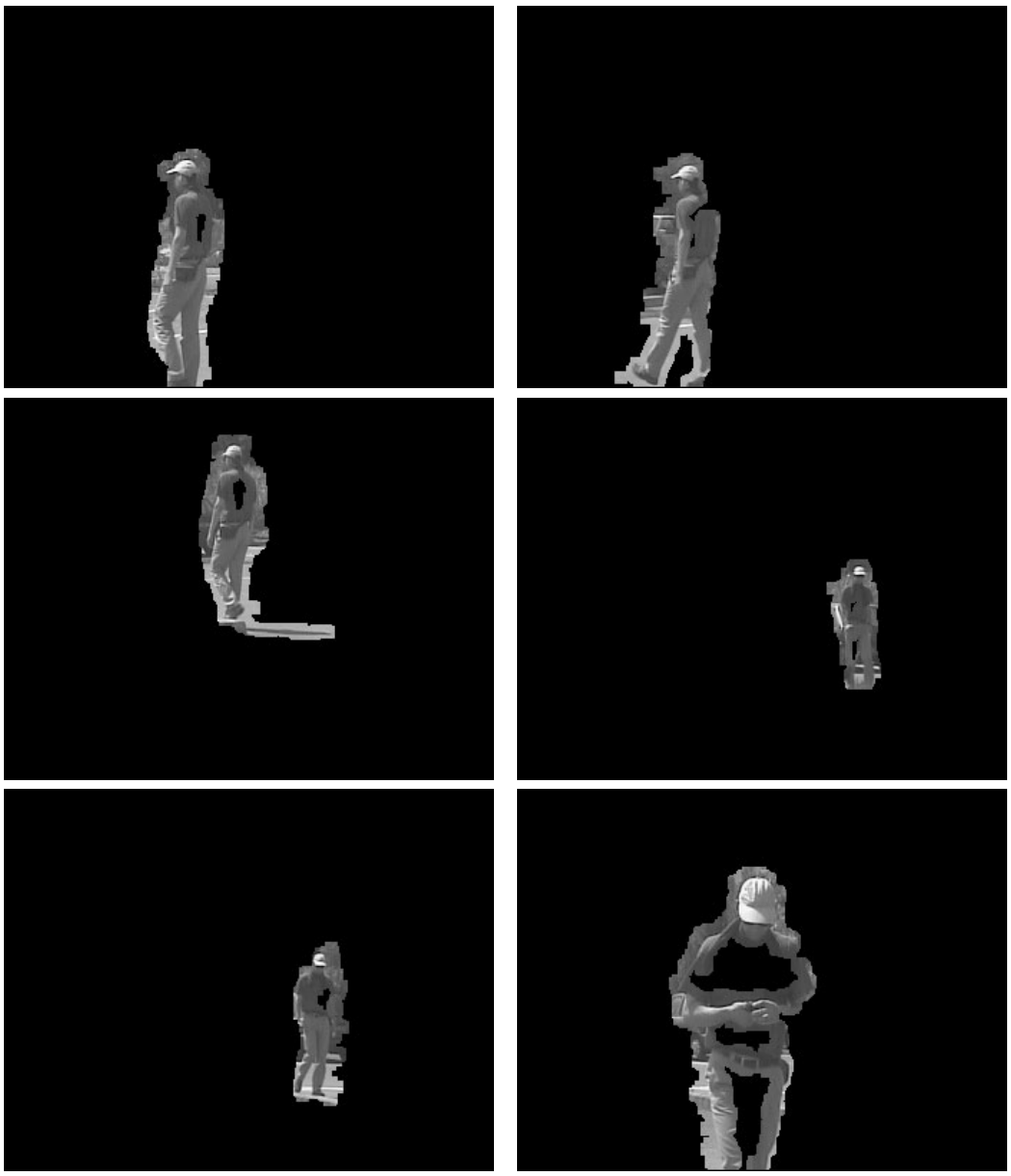

Figure 6: This figure shows the foreground pixels extracted from the sequence shown in the previous image.

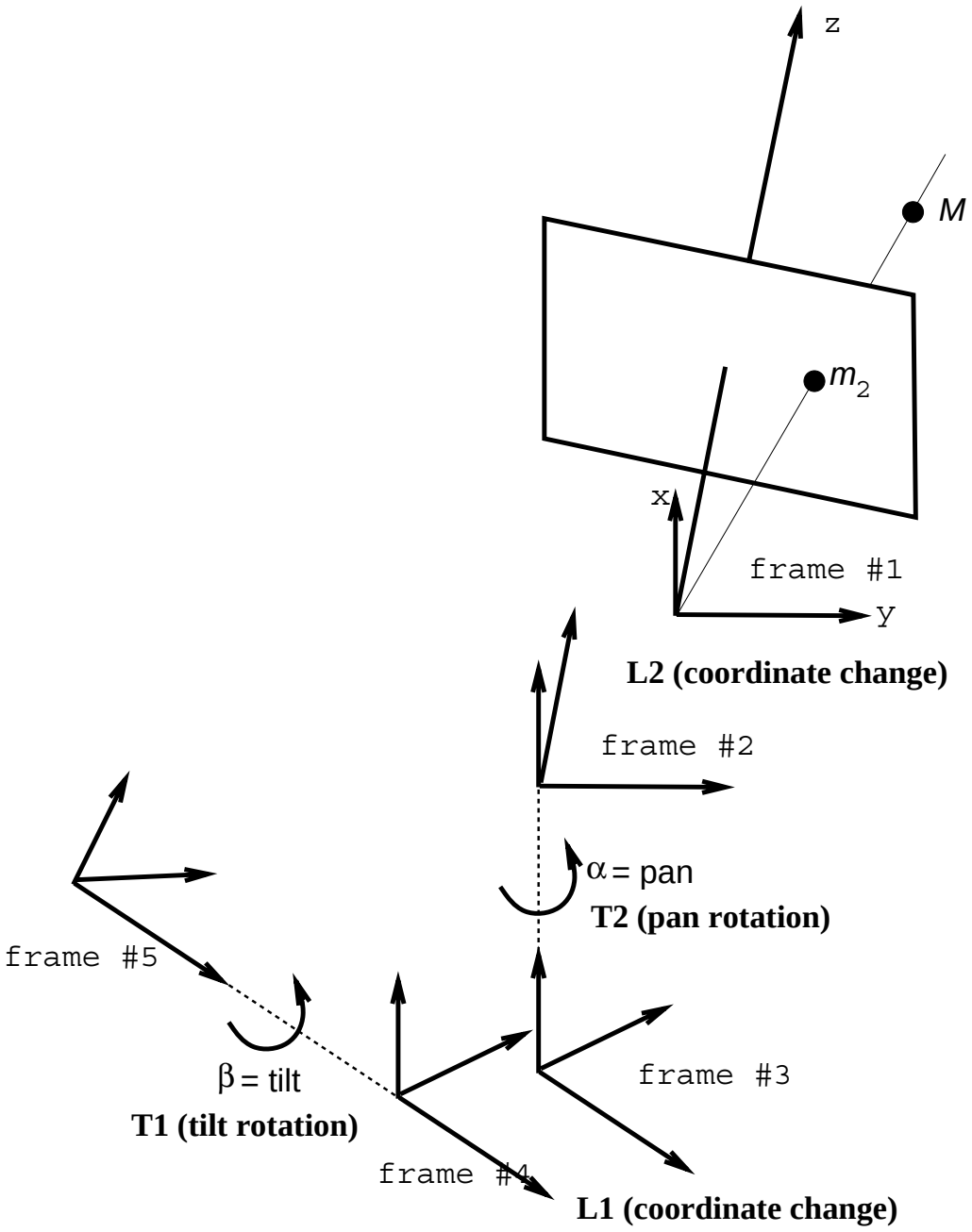

Figure 7: This figure shows a general pan-tilt mechanical model which attaches a camera (frame #1) to a rigid fixture (frame #5). Estimating the pan and tilt angles such that a given scene point appears at the image center is a non-trivial inverse kinematic problem.

Unité de recherche INRIA Rhône-Alpes
655, avenue de l'Europe - 38330 Montbonnot-St-Martin (France)

Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399