Camera Cooperation for Achieving Visual Attention

Radu Horaud, David Knossow, Markus Michaelis

Camera cooperation for achieving visual attention

Radu Horaud, David Knossow, and Markus Michaelis
INRIA Rhône-Alpes
655, avenue de l'Europe
38330 Montbonnot Saint-Martin, FRANCE
Corresponding author: Radu Horaud
Radu.Horaud@inrialpes.fr
fax: +33 476 615 454

April 27, 2011

Machine Vision and Applications,
16(6), pp 331–342, February 2006

Abstract

In this paper 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.

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

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 particular, 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 [15], [8], [6], [16]. 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 of multiple camera geometry.

In this paper we propose an innovative solution that combines the advantages of both static and active cameras and of both low- and high-resolution images. One camera is fixed and has a wide field of view, thus allowing surveillance of a wide area in terms of both width and depth of its 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 mounted onto a motor-driven pan and tilt device. 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 [18]. With respect to [18] which describes a general philosophy and a system architecture, we analyse and characterize in detail the fine geometric and mechanical coupling between a static and a rotating camera.

In practice, 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 the geometrical and the 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 original contributions. contributions: We derive a mathematical expression for the two-camera coupling, where one camera is static and the other comera rotates, under the form of a set of polynomial equations. We show that, in the general case, there may be several solutions for the pan and tilt angles and that these solutions are parameterized by the a depth parameter (the depth from the static camera to the scene event). We consider the special case where the pan and tilt rotational axes are mutually orthogonal. We show that with a practical camera setup there is a unique solution for the pan and tilt values. We describe a practical solution for achieving gaze control with a rotating camera and for separating a moving object from its static background. Once an initial solution is found, gaze-control is reduced to the tracking of an event in the static image and to the updating of the pan and tilt angle values.

The remainder of this paper is organized as follows.

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 analyse both the general case and a simplified pan-tilt model. We derive the number of algebraic solutions.

Section 3 *describes a method for dynamically separating an event (a moving object) from its background* by estimating the projective mapping associated with a camera undergoing rotational motions. 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.

Appendices A, B, and C provide a detailed description of the kinematic model being used to describe the pan and tilt device, as well as a method for calibrating the fixed parameters of this zero-reference kinematic model.

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*.

In this section we establish the formal link between a static camera and a rotating camera based on the epipolar geometry (which holds at each time instant) and the kinematic model associated with a pan and tilt mechanism. First we introduce the point reconstruction equations. Second we consider a pan and tilt kinematic model in its most general form. Third, we analyse the case of a *s*implified pan and tilt model, i.e., the pan and tilt rotation axes are mutually orthogoanal.

2.1 Two-camera geometry

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 in projective space by a 4-vector $\boldsymbol{M} = (X\ Y\ Z\ 1)^\top$, 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 with projective

coordinates $\boldsymbol{m}_1$ and $\boldsymbol{m}_2$. For pinhole cameras, the 3×4 projection matrices have the following parameterization:

$$\begin{aligned} \mathbf{P}_1 &= \mathbf{K}_1 \left[\begin{array}{cc} \mathbf{I} & \mathbf{0} \end{array} \right] \quad (3) \\ \mathbf{P}_2 &= \mathbf{K}_2 \left[\begin{array}{cc} \mathbf{R} & \boldsymbol{t} \end{array} \right] \quad (4) \end{aligned}$$

The 3×3 matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ have the intrinsic camera parameters as entries (see below the expression of $\mathbf{K}_2$). 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 may or may not be known and which is allowed to vary. The expression of $\mathbf{K}_2$ is:

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

In order to eliminate the known camera parameters from the equations we use the the substitutions $\boldsymbol{m}_1 = \mathbf{K}_1 \boldsymbol{n}_1$ and $\boldsymbol{m}_2 = \mathbf{K}'_2 \boldsymbol{n}_2$. By combining eq. (1) with eq. (3) we obtain a simple expression for the coordinates of $\boldsymbol{M}$:

$$\boldsymbol{M} = \begin{pmatrix} \lambda_1 \boldsymbol{n}_1 \\ 1 \end{pmatrix}$$

By combining eq. (2) with eq. (4) and by substitution of $\boldsymbol{M}$ we obtain:

$$\lambda_2 \boldsymbol{n}_2 = \mathbf{D}_f \left(\lambda_1 \mathbf{R} \boldsymbol{n}_1 + \boldsymbol{t} \right) \quad (5)$$

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 the notation $\boldsymbol{n}_2 = (x_2 \; y_2 \; 1)^\top$ we farther eliminate λ_1 by dividing the first and second vector components, $()_1$, $()_2$ with the third vector component, $()_3$:

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

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:

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

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 (i) the relative position of the active camera with respect to the static camera and of (ii) the kinematic model associated with the active camera's pan and tilt mechanism. Therefore we must establish the link between the epipolar geometry constraint and the kinematic model constrains. 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 and constrained motion* undergone by the active camera from its docking position to a current position. From Figure 1 one can notice that the following relationship holds:

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

Insert Figure 1 approximatively here

2.2 General pan-tilt model

Matrices $\mathbf{Q}$ and $\mathbf{T}_0$ have the same mathematical structure although the former describes a *kinematically constrained motion* while the latter describes a *static* relationship between two Cartesian frames. Matrix $\mathbf{Q}$ describes the motion undergone by a pan and tilt mechanism. In order to describe such a mechanism we will adopt the well known zero-reference kinematic model. The latter is described in many textbooks such as [13, 17, 14]. 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: 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:

$$\begin{aligned} \mathbf{R}_1 &= \mathbf{I}_{3\times 3} + \sin(\beta - \beta_0)\hat{\mathbf{R}}_1 + (1 - \cos(\beta - \beta_0))\hat{\mathbf{R}}_1^2 & (15)\\ \boldsymbol{t}_1 &= \sin(\beta - \beta_0)\hat{\boldsymbol{t}}_1 + (1 - \cos(\beta - \beta_0))\hat{\mathbf{R}}_1\hat{\boldsymbol{t}}_1 & (16) \end{aligned}$$

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

$$\begin{aligned} \mathbf{R} &= \mathbf{R}_2\mathbf{R}_1\mathbf{R}_0 & (17)\\ \boldsymbol{t} &= \mathbf{R}_2\mathbf{R}_1\boldsymbol{t}_0 + \mathbf{R}_2\boldsymbol{t}_1 + \boldsymbol{t}_2 & (18) \end{aligned}$$

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

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

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:

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

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. From the equations above it can

be observed that $\mathbf{R}_1$, $\boldsymbol{t}_1$, $\mathbf{R}_2$, and $\boldsymbol{t}_2$ are parameterized by the known tangent operators (see appendix C) and by the three unknowns – the pan and tilt angles and the depth parameter λ_2. 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 standard substitutions:

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

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

we obtain three polynomial equations in three unknowns: t_α, t_β, and λ. It is possible to eliminate λ as an unknown between the second and third equations, at the price of increasing the degree of the resulting polynomials. 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.

2.3 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. This simpler kinematic model allows an algebraic analysis of the number of solutions associated with the inverse kinematics of the pan and tilt camera. Moreover and for the sake of this analysis, 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 may 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.

In the past, event background separation has been mainly addressed with static cameras. When a camera moves, the problem is more difficult because one has to distinguish between camera motion (egomotion) and event motion. Nevertheless, whenever the camera undergoes a *pure rotational motion*, i.e., when the center of projection lies onto the axis of rotation, it is possible to separate egomotion from event motion by assuming that the background pixels are transformed from one image to another by a 2-D projective mapping, [10].

The motion of the pan and tilt camera is described by eq. (10). In general, this does not guarantee that the camera undergoes a pure rotation around its center of projection because of the mechanical offsets. In practice the latters are *small* compared to the distance from the camera to the background and therefore the background may well be viewed as a plane at infinity, [10].

Let $\boldsymbol{m}_2^{t-1}$ and $\boldsymbol{m}_2^t$ describe the homogeneous coordinates of an image point at times $t-1$ and t. The subscript $_2$ indicates that we deal with the active camera. One can apply equations (3) and (4) to the active camera and assume that the translational part of the motion is null. We obtain the following well known formula for cameras undergoing pure rotation:

$$\boldsymbol{m}_2^t = \mathbf{K}_2 \mathbf{R}_2^{t,t-1} \mathbf{R}_1^{t,t-1} \mathbf{K}_2^{-1} \boldsymbol{m}_2^{t-1} \tag{23}$$

where $\mathbf{R}_2^{t,t-1}\mathbf{R}_1^{t,t-1}$ models the rotation of the active camera. We denote this mapping by:

$$\mathbf{H}^{t,t-1} = \mathbf{K}_2\mathbf{R}_2^{t,t-1}\mathbf{R}_1^{t,t-1}\mathbf{K}_2^{-1} \tag{24}$$

and the problem is to estimate the 3×3 matrices H as the camera rotates.

The relationship between $\boldsymbol{m}_2^{t-1}$ and $\boldsymbol{m}_2^t$ 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 two-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.

Insert Figure 2 approximatively here

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 (for the sake of simplicity we drop the subscript $_2$):

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

The function $\Psi()$ denotes the non linear mapping from homogeneous to Euclidean coordinates of $\boldsymbol{m}$, $\Psi(m_1, m_2, m_3) = (m_1/m_3,\ m_2/m_3)^\top$. Various methods were developed in the past for solving this non-linear minimization problem [11], [21], [1].

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 football player, onto a bicycle, onto a pedestrian, or onto 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 nearby its image center, even if the object's appearance changes, if its depth changes, and/or if the object is partially occluded. This process requires three steps: off-line calibration, initialization and gaze control.

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

- Off-line calibration: see section 4.1.
- Initialization:
  - A scene object is detected and tracked over time (automatically, semi-automatically, or manually) using the static camera;
  - The active camera rotates such that this scene object falls within its field of view and the depth of this scene object is estimated, i.e., section 4.2;
  - Pan and tilt values are estimated *from scratch* by solving a set of three polynomials associated with the inverse kinematics of the pan-and-tilt device and the active camera is rotated accordingly;
- Gaze control:
  - The pan and tilt angle values estimated at time $t-1$ are used as initial guesses to find their values at time t. Notice that the depth information is maintained constant and the consequences of this choice are explained below.
  - Images at times $t-1$, t, and $t+1$ are used to separate the moving object from the background.

Notice that during the gaze-control stage of the algorithm the depth associated with the scene object is not updated: Instead, its previously estimated value (during the initialization stage) is used. As a consequence, the object will not appear at the image center of the active camera as its depth varies over time.

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 [19].

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 [20] 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 2.2 and 2.3 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.

Insert Figure 3 approximatively here

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

A full set of experiments is summarized through Figures 4, 5, 6, 7, 8, and 9. 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 8 samples out of a 550-frame image sequence.

In the first example (figures 4, 5, 6) the object of interest is a pedestrian. During the initialization phase, this object 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 depth from the pedestrian to the static camera.

Insert Figure 4 approximatively here
Insert Figure 5 approximatively here
Insert Figure 6 approximatively here

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. Notice however that the pedestrian is not displayed at the image center. This is because there is an error in the depth estimate. The pan and tilt values are computed based on a depth estimation that is different than the true depth value.

It is worthwhile to notice the behaviour of the system in the presence of occlusions and of missing data. The pedestrian is first occluded by a car, then appears and then walks outside the field of view of the camera, turns, and comes back. Instead of these disturbances the gaze of the active camera is correctly controlled. Whenever the object disappears from the field of view of the static camera, the active camera tracks the moving object using the event/background separation method outlined above.

In the second example (figures 7, 8, 9) the object of interest is a bicycle rider. Notice that the object is properly tracked in spite of partial occlusions by surrounding objects. In order to assess the quality of homography estimation between consecutive images in the sequence, we removed the foreground pixels and built a "foreground" image sequence, as shown in Figure 9.

Insert Figure 7 approximatively here
Insert Figure 8 approximatively here
Insert Figure 9 approximatively here

From a more practical point of view, the size of the images is 640×480. The focal length of the static camera is of 500 pixels. Event/background separation operates on 320×240 images. The whole two-camera system runs at 10 frames per second on a 1.7GHz processor.

5 Conclusion

In this paper we addressed the problem of coupling two cameras in order to achieve visual attention – controlling a camera to gaze in a selected direction. 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 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 camera and a rotating camera. We derived a solution for this coupling both for a general kinematic mechanism and for a simpler pan-tilt model. We showed that under the practical setup that we used, 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 gaze-control loop is activated in order to estimate the camera's rotational degrees of freedom. Moreover, the system is able to use event detection (performed with the static camera) in order to facilitate event/background segmentation performed with a rotating camera.

The camera cooperation principle developed in this paper could easily be generalized to several rotating cameras. Therefore, multiple moving objects detected with the static camera could be handled separately by multiple rotating cameras.

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 10:

Insert Figure 10 approximatively here

- 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:

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

Revised paragraph

By eliminating $\boldsymbol{M}^{(5)}$ in between these two equations and by properly adding some *dummy* transformations, we obtain:

$$\begin{aligned}\boldsymbol{M}^{(1)}(\alpha,\beta) &= \mathbf{L}_2\mathbf{T}_2(\alpha)\mathbf{T}_2^{-1}(\alpha_0)\mathbf{L}_2^{-1}\mathbf{L}_2\mathbf{T}_2(\alpha_0)\mathbf{L}_1\mathbf{T}_1(\beta)\mathbf{T}_1^{-1}(\beta_0)\mathbf{L}_1^{-1}\mathbf{T}_2^{-1}(\alpha_0)\mathbf{L}_2^{-1}\boldsymbol{M}^{(1)}(\alpha_0,\beta_0)\\ &= \underbrace{\mathbf{L}_2\mathbf{T}_2(\alpha-\alpha_0)\mathbf{L}_2^{-1}}_{\mathbf{Q}_2}\underbrace{\mathbf{L}_2\mathbf{T}_2(\alpha_0)\mathbf{L}_1\mathbf{T}_1(\beta-\beta_0)\mathbf{L}_1^{-1}\mathbf{T}_2(-\alpha_0)\mathbf{L}_2^{-1}}_{\mathbf{Q}_1}\boldsymbol{M}^{(1)}(\alpha_0,\beta_0)\end{aligned}$$

This is 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 $\mathbf{L}_2$ (see appendix A and Figure 10) 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. *Computer Animation and Virtual Worlds*, 15(6):501–517, November 2004.

[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.

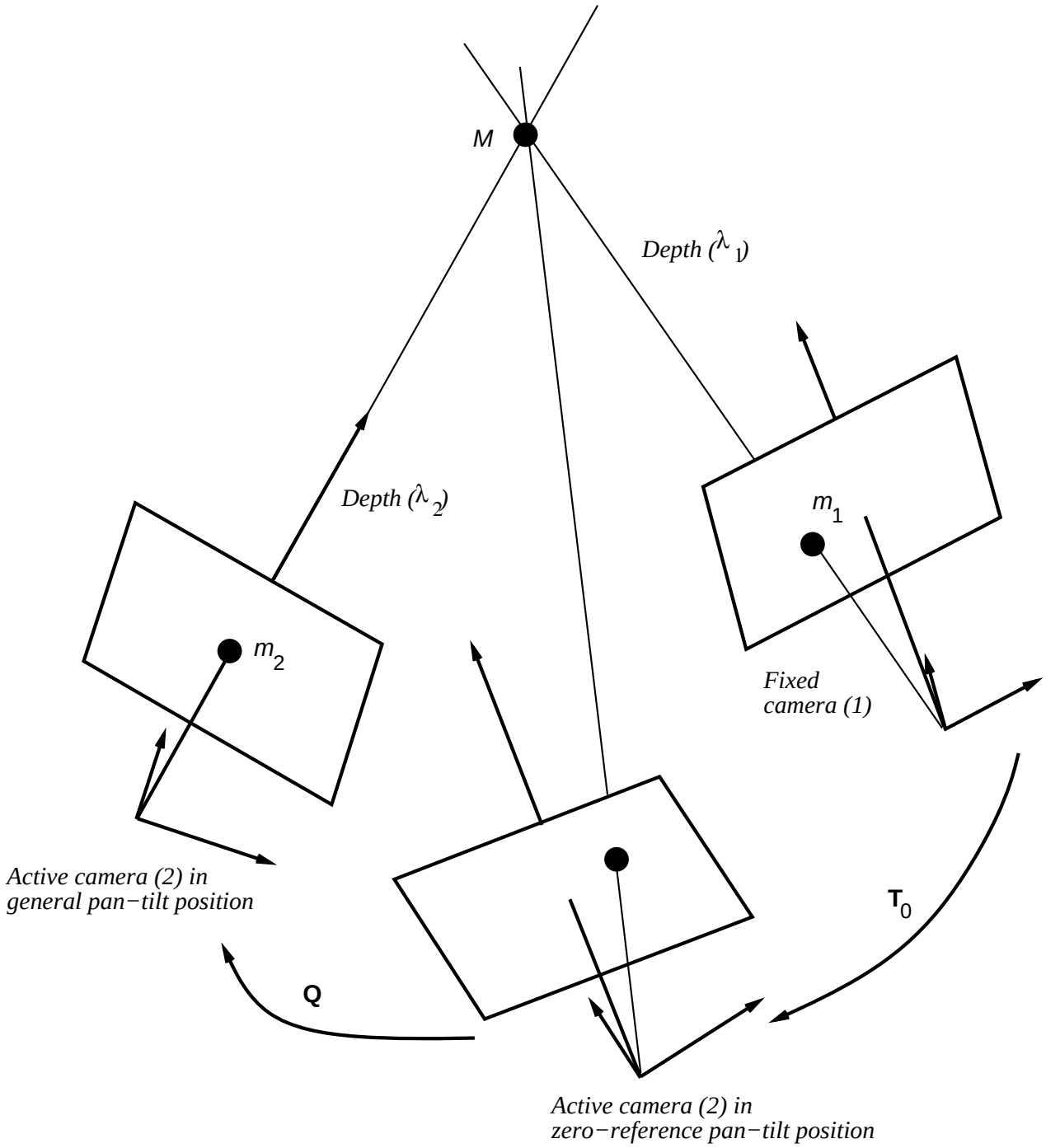

Figure 1: The active camera has a docking or a zero-reference position. Both the stereo (matrix $\mathbf{T}_0$) and kinematic (matrix $\mathbf{Q}$) calibrations are performed with respect to this reference position.

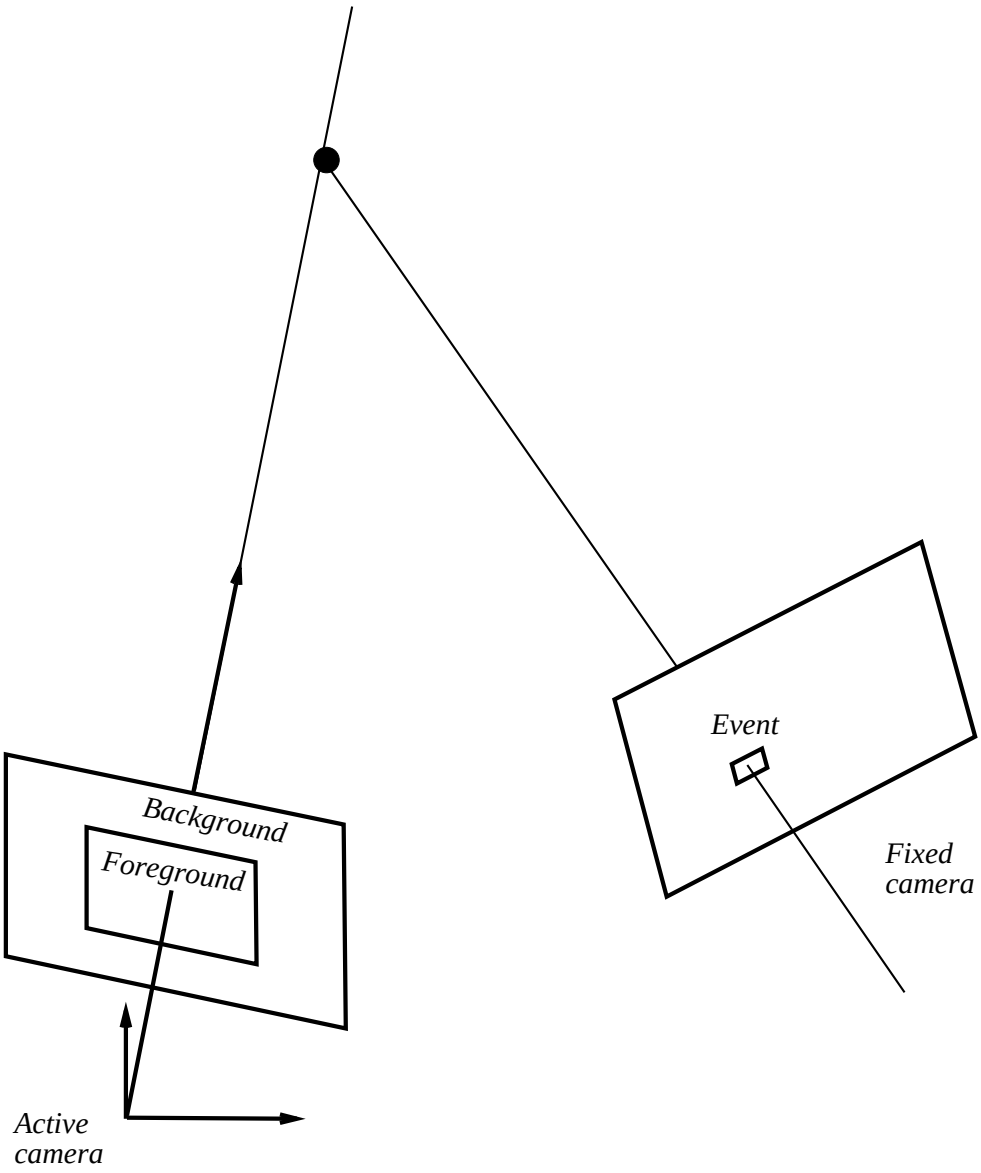

Figure 2: The coupling between the cameras allows one to associate *foreground* and *background* regions with the active camera's image. The event, which is predicted in the static camera at low resolution, must lie in the foreground region associated with the active camera's image sequence.

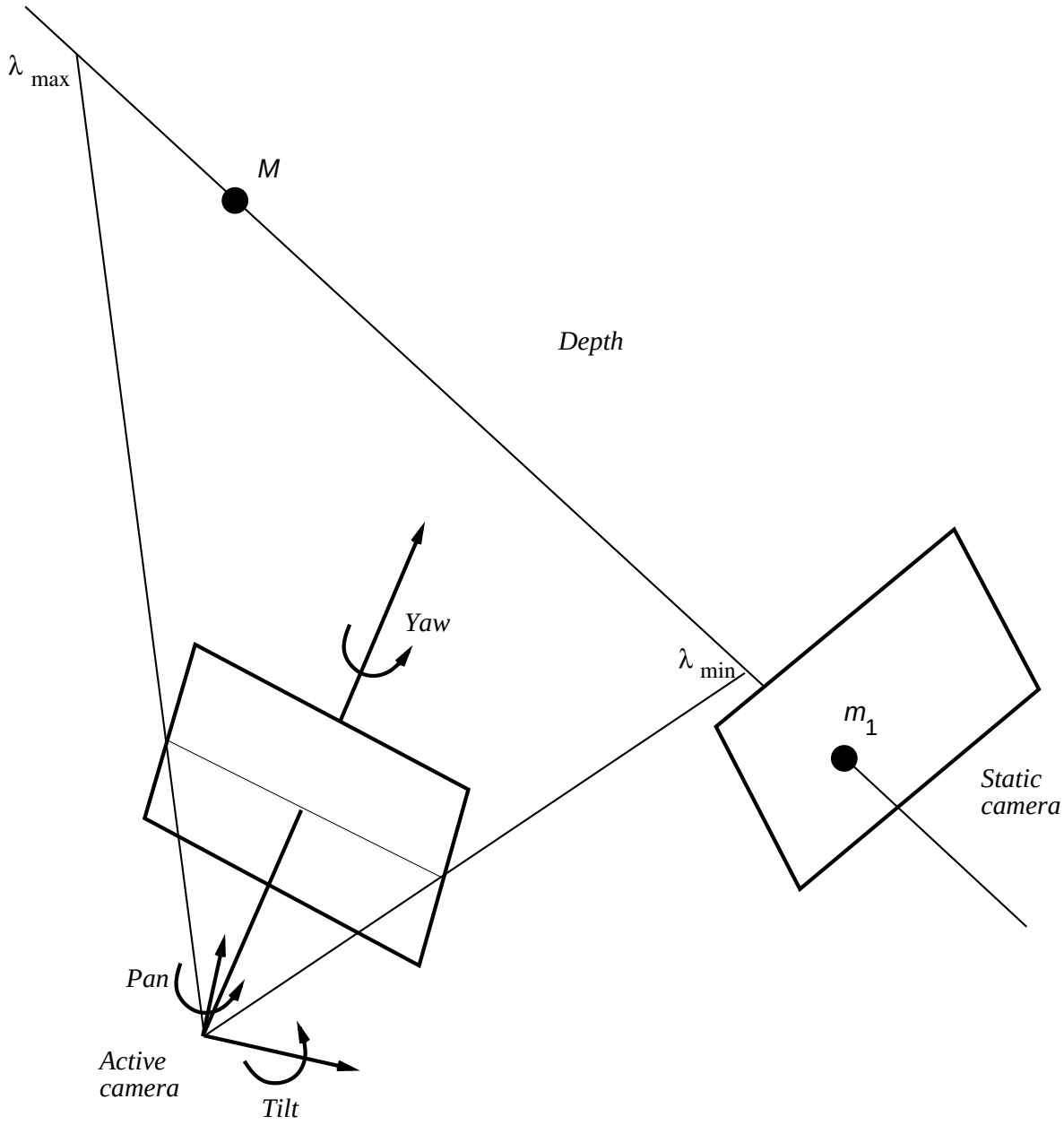

Figure 3: In order to estimate the depth to the point M, the active camera must see this point. The degrees of freedom of the active camera – pan, tilt, yaw, and focal length – are estimated such that the line of sight associated with image point m_1 is seen in the active image plane.

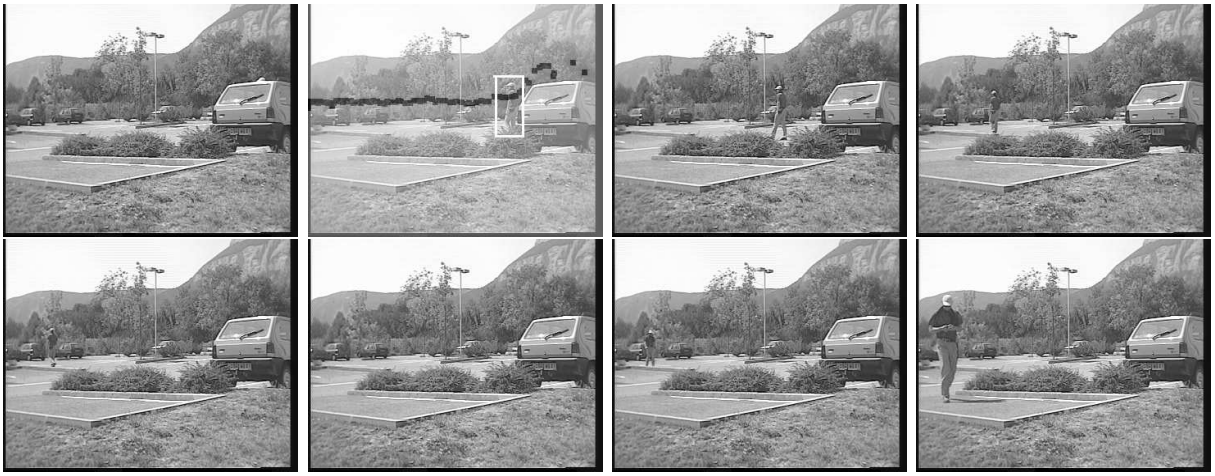

Figure 4: This figures shows 8 frames out of a 550-frame image sequence available with the static camera. The second frame shows the trajectory of the moving person.

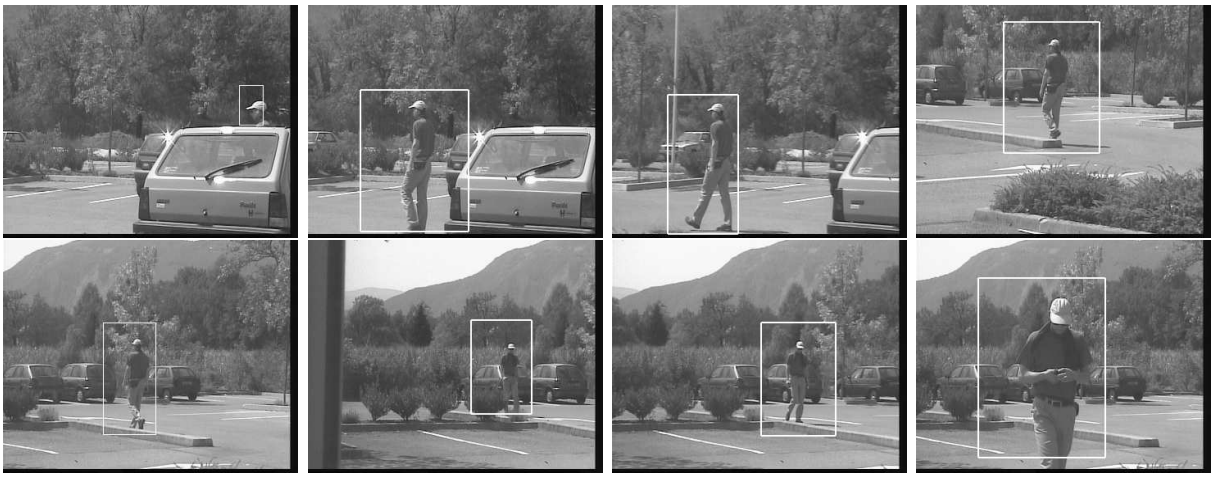

Figure 5: The output of the active camera after gaze control.

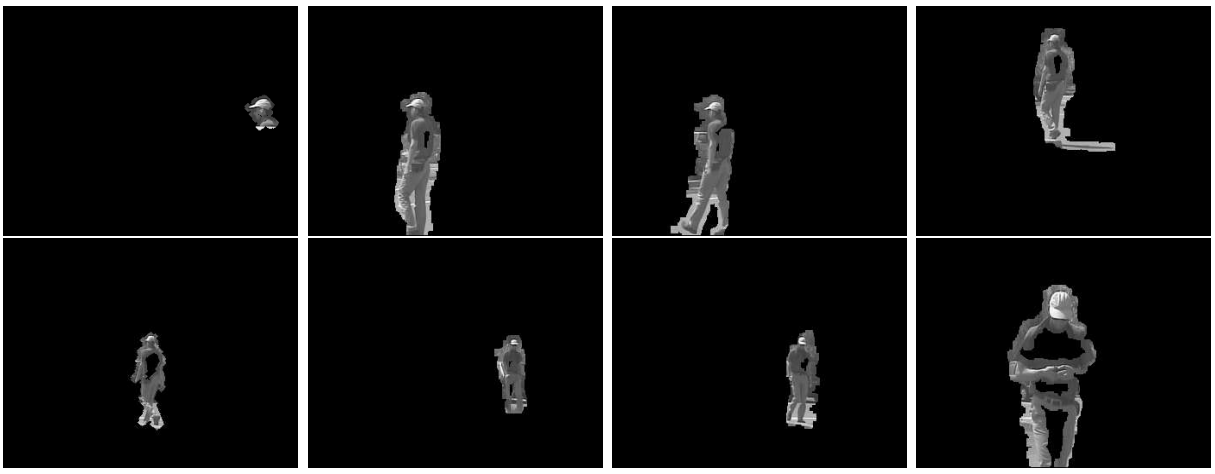

Figure 6: The foreground pixels extracted from the sequence shown above.

Figure 7: A second example showing 8 frames associated with the active camera.

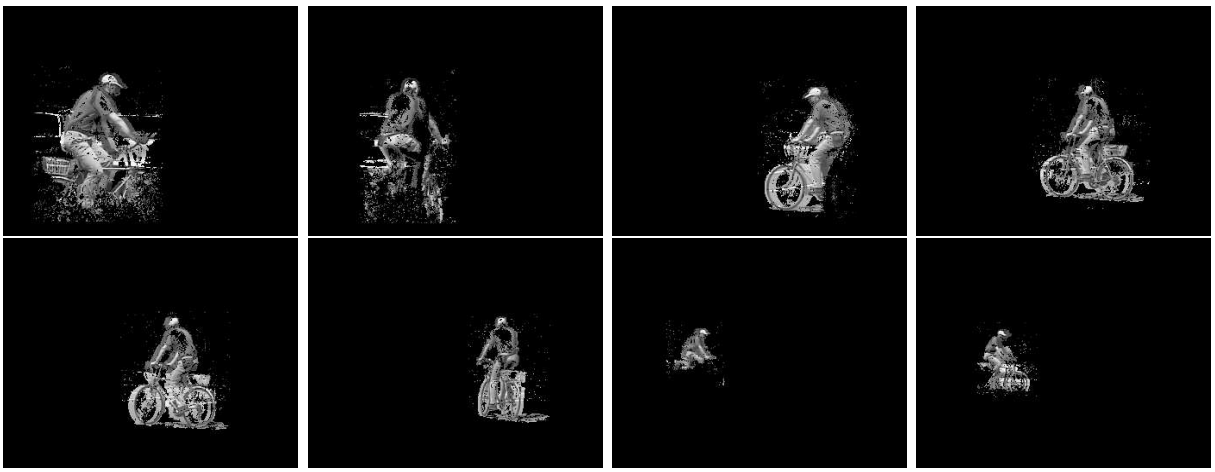

Figure 8: The result of foreground detection applied to the second example

Figure 9: The foreground pixels were removed from the image sequence and replaced by background using the image-to-image homographies.

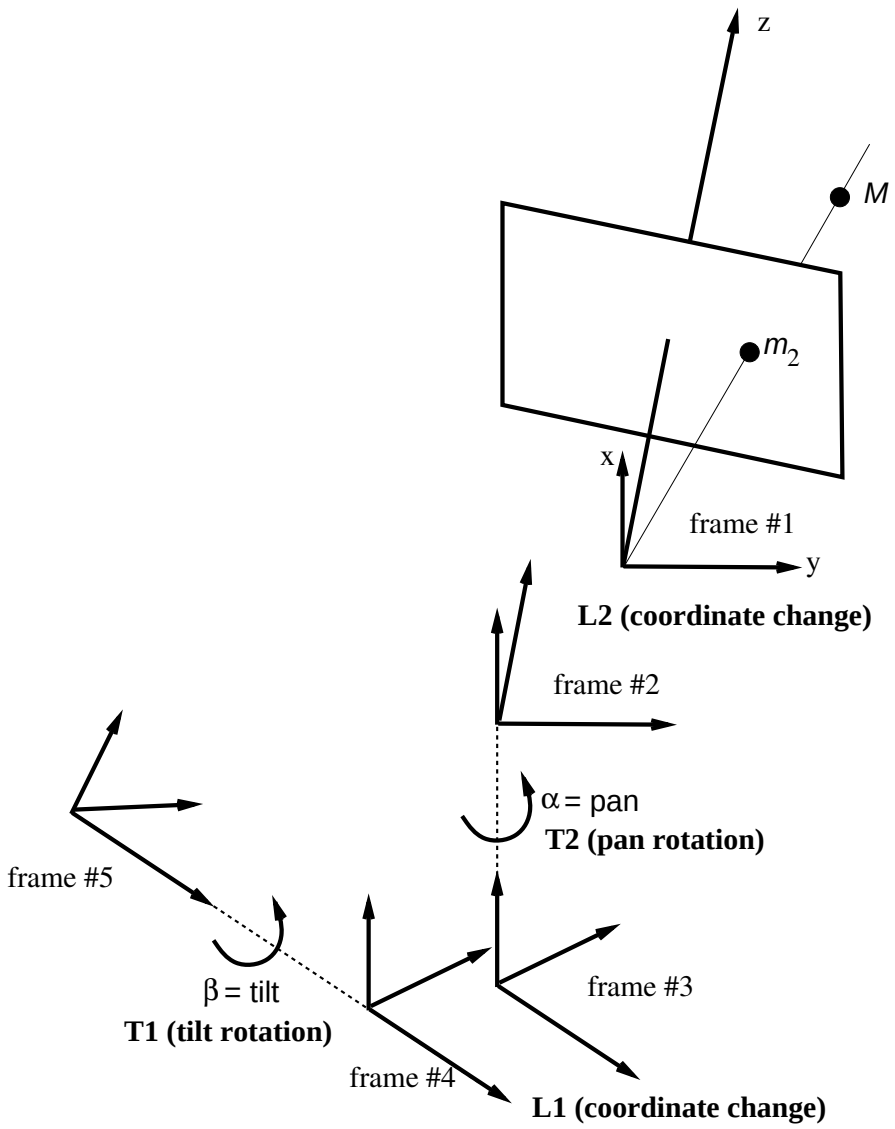

Figure 10: This figure shows a general pan-tilt mechanical model which attaches a camera (frame #1) to a rigid fixture (frame #5). Estimating the pan and tilt angles such that a given scene point appears at the image center is a non-trivial inverse kinematic problem.

[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] B. W. Mooring, Roth Z. S., and M. R. Driels. *Fundamentals of Manipulator Calibration*. John Wiley & Sons, 1991.

[15] 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.

[16] 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.

[17] R.M. Murray, Z. Li, and S.S. Sastry. *A Mathematical Introduction to Robotic Manipulation*. CRC Press, Ann Arbor, 1994.

[18] 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.

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

[20] 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.

[21] 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.