Reactive object tracking with a single PTZ camera

Murad Al Haj, Andrew D. Bagdanov, Jordi González and F. Xavier Roca
Departament de Ciències de la Computació and Centre de Visió per Computador,
Universitat Autònoma de Barcelona, 08193, Bellaterra, Spain
{malhaj, bagdanov}@cvc.uab.es {jordi.gonzalez, xavier.roca}@uab.es

Abstract

In this paper we describe a novel approach to reactive tracking of moving targets with a pan-tilt-zoom camera. The approach uses an extended Kalman filter to jointly track the object position in the real world, its velocity in 3D and the camera intrinsics, in addition to the rate of change of these parameters. The filter outputs are used as inputs to PID controllers which continuously adjust the camera motion in order to reactively track the object at a constant image velocity while simultaneously maintaining a desirable target scale in the image plane. We provide experimental results on simulated and real tracking sequences to show how our tracker is able to accurately estimate both 3D object position and camera intrinsics with very high precision over a wide range of focal lengths.

1. Introduction

Many applications in computer vision benefit from high resolution imagery, for example license-plate identification or face recognition where a minimum size of the target is required. The problem with zoom control is that two opposing goals are desirable: to obtain maximum resolution of the tracked target while simultaneously minimizing the risk of losing it. With a finite number of fixed sensors, there is a fundamental limit on the total area that can be observed. Maximizing both the area of coverage and the resolution of each observed target requires an increase in the number of cameras. Such multi-camera systems are both cumbersome and costly. Therefore, a system utilizing a single pan-tilt-zoom (PTZ) camera can be much more efficient if it is properly designed to overcome the obvious drawback of having less sensors viewing the scene.

Towards this end, different works have investigated the use of PTZ cameras to address this problem of actively surveying a large area in an attempt to obtain high-quality imagery while maintaining coverage of the region [6]. Starting two decades ago, the area of active vision has been gaining much attention, in an attempt to improve the quality of the acquired visual data by keeping a certain object at a constant scale. Early on, Aloi et al., introduced the first general framework for active vision in order to improve the perceptual quality of tracking results [2]. Denzler et al. modeled the motion of the tracked object using a Kalman filter trying to select the camera focal length that minimizes the uncertainty in state estimation with respect to the observation [4]. However, the authors used a dual-camera, stereo set-up which significantly simplified the 3D estimation problem. A newer approach is described by Tordoff et al., which tunes a constant velocity Kalman filter in order to insure reactive zoom tracking while the focal length is varying [7]. Their approach is to correlate all the parameters of the filter with the focal length. However, they don’t concentrate on the overall estimation problem, and their filter does not take into account any real-world object properties. A distinguishing characteristic of our approach is that we incorporate depth estimation directly in our filter. The most recent work is that of Neslon et al., where a second camera with fixed focal length is introduced in order to solve the problem of lost fixation [5]. The focus of the latter two works is primarily on zoom control so they do not deal with total object-camera position estimation and its use in the control process. An attempt to join estimation and control in the same framework can be found in the work of Bagdanov et al., where a PTZ camera is used to actively track faces [3]. However, both the estimation and control models used are ad hoc, and the estimation approach used is based on image features rather than 3D properties of the target being tracked.

Accurate reactive tracking of moving objects is a problem of both control and estimation. The speed at which the camera is adjusted must be a joint function of the current camera position in pan, tilt and focal length, in addition to the position of the tracked object in the 3D
world. In this paper we formulate the problem of jointly estimating the camera state and the 3D object position as a Bayesian estimation problem. In section 2, we describe the joint model of the camera and the 3D world, providing the stage for estimation and control, which is formulated in section 3. In section 4, we report on experiments conducted on both simulated data and live cameras. We conclude in section 5 with a summary and indications of future research directions.

2. The camera and world model

In our method, a pinhole camera model is used. The camera center is located at the origin of the world coordinate system, the principal point is at the origin of the plane of projection and at zero pan and tilt. The axis of projection is aligned with the z-axis and the center of the object.

The object being tracked is assumed to be a planar patch with known width \( W \) and height \( H \), located at world position \( (X,Y,Z)\)\(^T\). Changes in camera orientation due to panning and tilting are modeled as pure rotations of the coordinate system:

\[
R(T,P) = \begin{bmatrix}
1 & 0 & 0 \\
0 & \cos T & -\sin T \\
0 & \sin T & \cos T
\end{bmatrix},
\]

where \( P \) and \( T \) represent the pan and tilt angles, respectively.

We assume that the camera projection is reasonably approximated using equal scaling in the \( x \) and \( y \) directions on the plane of projection (i.e. square pixels). The center of projection is also assumed to be at the origin of the world coordinate system. The camera matrix is then parametrized by a single focal length parameter \( F \):

\[
K(F) = \begin{bmatrix}
F & 0 & 0 \\
0 & F & 0 \\
0 & 0 & 1
\end{bmatrix}.
\]

The projection of the object at position \( O = (X,Y,Z)\)\(^T\) onto the plane of projection can now be written as:

\[
f(P,T,F,O) = \begin{bmatrix}
X' \\
Y' \\
Z'
\end{bmatrix},
\]

where \( X' \), \( Y' \) and \( Z' \) are given by the transformation:

\[
\begin{bmatrix}
X' \\
Y' \\
Z'
\end{bmatrix} = K(F)R(T,P)O.
\]

This camera model relates the geometry and position of the tracked object in the 3D world to the internal camera parameters. In the next section we describe how the estimation problem can be formulated.

3. Estimation and control

In this section we formulate the problem of jointly estimating the camera and world parameters in a recursive Bayesian filter framework. At time \( k \), the state configuration of the joint camera/object model is represented by the spatial coordinates of the tracked object in the real world, the camera intrinsics and the velocities corresponding to the object position and camera intrinsics:

\[
x_k = [O_k \mid C_k \mid \dot{O}_k \mid \dot{C}_k]^\top,
\]

where each component is defined as:

\[
O_k = [X_k, Y_k, Z_k], \quad C_k = [P_k, T_k, F_k],
\]

\[
\dot{O}_k = [\dot{X}_k, \dot{Y}_k, \dot{Z}_k], \quad \dot{C}_k = [\dot{P}_k, \dot{T}_k, \dot{F}_k].
\]

\([X_k, Y_k, Z_k]\) is the position of the planar patch in world coordinates at time \( k \), and \([P_k, T_k, F_k]\) represent the camera pan angle, tilt angle and focal length, respectively. The remaining elements, \([\dot{X}_k, \dot{Y}_k, \dot{Z}_k, \dot{P}_k, \dot{T}_k, \dot{F}_k]\), represent the velocities of the previously mentioned components.

From time \( k-1 \) to time \( k \), the state is updated by the linear function \( U \):

\[
x_k = U x_{k-1} + v_{k-1},
\]

where \( U \) is defined as:

\[
U = \begin{bmatrix}
I_6 & I_6 \\
0_6 & I_6
\end{bmatrix},
\]

where \( I_n \) and \( 0_n \) are the \( n \times n \) identity and zero matrices, respectively. The term \( v_{k-1} \) in equation (2) is considered to be a zero-mean, Gaussian random variable adding noise to the system update.

At each time \( k \), an observation \( z_k \) of the unknown system \( x_k \) is made. \( z_k \) is defined as:

\[
z_k = [x^p_k, y^p_k, w^p_k, h^p_k, p_k, t_k, f_k],
\]

where \((x^p_k, y^p_k)\) is the center of the object in the image plane measured in pixels, \((w^p_k, h^p_k)\) are the width and height of the object in the image plane, also measured in pixels and \((p_k, t_k, f_k)\) are the camera parameters arriving from the camera imprecise measurements of the pan angle, tilt angle and focal length. The measurement equation, against which the observation \( z_k \) is compared, is given by:

\[
h_k(x_k) = [o_k \mid c_k]^\top + [n^o_k \mid n^c_k]^\top,
\]

where \( n^o_k \) and \( n^c_k \) are zero-mean Gaussian processes on the object and camera measurements, respectively, and \( o_k \) and \( c_k \) are defined as:

\[
o_k = [f(P_k, T_k, F_k, O_k) \mid f(0, 0, F_k, [W, H, Z_k]^\top)]^\top,
\]

\[
c_k = [P_k, T_k, F_k].
\]
where $Z'_k$ is the projection of the depth $Z_k$ in the new coordinate system resulting from the pan and tilt of the camera. The object measurement $\mathbf{o}_k$ consists of the projection of the object position $\mathbf{O}_k$ and the known object size $W \times H$ into the coordinate system of the image plane. The camera measurement $\mathbf{c}_k$ represents the pan angle, tilt angle and focal length, as estimated by the state vector.

Given the system update and measurement processes defined in equations (2) and (4), the Bayesian estimation problem is to find an estimate of the unknown state $\mathbf{x}_k$ that maximizes the posterior density $p(\mathbf{x}_k|z_{1:k})$. Towards this end, an Extended Kalman filter (EKF) is implemented to recursively solve this estimation problem [8].

The estimated state output at each step of the Kalman filter is used to control the movement of the camera. Two PID controllers are used: one for controlling the pan and tilt and another for the zoom. The control signal, outputted by a PID controller, is given by:

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de(t)}{dt},$$

where $e(t)$ is the error signal. In our case, and at each time $t$, the error in pan is defined as the difference between the estimated pan angle and the estimated horizontal angle the object forms with the world coordinate system, while the error in tilt is defined as the difference between the estimated tilt angle and the estimated vertical angle of the object:

$$e_p = \arctan(X_k/Z_k) - P_k,$$

$$e_t = \arctan(Y_k/Z_k) - T_k.$$

The gains are experimentally set to: $K_p = 1$, $K_i = 0$ and $K_d = 0.2$.

To calculate the error for the zoom controller, we define the desired area $D_a$, which is the maximum area in pixels we aim to have and which is usually achieved when the object is static. The error is then defined as:

$$e_z = D_a - w_{proj} \times h_{proj},$$

where $w_{proj}$ and $h_{proj}$ are the projections of the width (W) and height (H) of the object in the image plane. The gains are experimentally set to: $K_p = 0.01$, $K_i = 0$ and $K_d = 0$. The integral phase was bypassed in both controllers, by setting $K_i$ to 0, because the output of the filter was found to be very accurate at steady state, i.e. when the object is centered with maximum zoom.

4. Experimental analysis

Experiments were done using both simulated scenarios and live scenes of a PTZ camera. The simulated scenario consisted of a random motion of an object whose size is 10x10 cm, and the error was averaged over many runs. The camera used in the live scenes was an Axis 214 PTZ network camera.

4.1 Results on simulated data

To validate our model, we performed a number of simulations. The error metric we used in all model parameters estimation is:

$$\text{RMSD}(\phi_i) = \sqrt{E((\phi_i - \hat{\phi}_i)^2)},$$

where $\phi_i$ is one of the model parameters defined in equation 1, $\hat{\phi}_i$ is the estimated model parameter and the expectation is taken over the entire sequence. The RMSD is measured for several runs of the simulation (we used 100 runs in our experiments), and the average RMSD is used as a measure of estimation performance.

Figure 1 shows a box-and-whisker summary of the RMSD for a simulation where a moving object is tracked by a moving camera. In these experiments we simulate the motion the camera would execute due to corrections coming from the PID controllers described in section 3. In these simulations, some noise is introduced in the different state parameters. To investigate sensitivity to varying measurement noise, this value is scaled by a constant $a \in \{1, 5, 10\}$. Similar results can be seen in figure 2 for camera parameters estimation. From these figures, one can conclude that scaling the uncertainty, by $a = 5$ and $a = 10$, predictably scales the RMSD error as well as the spread (most notably in $Z$ and $t$) and increases outliers. However, even with such increase, the estimates of both, the object position and the camera parameters, are very good.

4.2 Results on live cameras

We also tested our method by actively tracking objects using a live, commodity PTZ camera (Axis 214). Simple assumptions about object size were made, the
5. Conclusions and future work

In this paper we described an approach to active camera tracking, using a single pan-tilt-zoom camera, that jointly estimates the orientation and focal length of the camera and the position of the tracked object relative to the camera center in the 3D environment. Experiments show that the estimates are robust in the presence of camera motion and increased measurement noise. We are currently investigating the use of the particle filter on the same estimation problem, and we are looking at alternate methods of estimating Gaussian uncertainty projected into the image plane in order to improve filtering results.

Acknowledgements

This work has been supported by the Spanish Research Programs Consolider-Ingenio 2010: MIPRCV (CSD200700018) and Avanza I+D VicCoMo (TSI-020400-2009-133); along with the Spanish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02. Also, Murad Al Haj gratefully acknowledges the financial support from the Generalitat de Catalunya through an AGAUR FI predoctoral grant (IUE/2658/2007).

References