Information Theoretic Focal Length Selection for Real–Time Active 3–D Object Tracking

J. Denzler, M. Zobel*, H. Niemann

Chair for Pattern Recognition
University of Erlangen–Nuremberg
91058 Erlangen, Germany
email: {denzler,zobel,niemann}@informatik.uni-erlangen.de

Abstract

Active object tracking, for example, in surveillance tasks, becomes more and more important these days. Besides the tracking algorithms themselves methodologies have to be developed for reasonable active control of the degrees of freedom of all involved cameras.

In this paper we present an information theoretic approach that allows the optimal selection of the focal lengths of two cameras during active 3–D object tracking. The selection is based on the uncertainty in the 3–D estimation. This allows us to resolve the trade–off between small and large focal length: in the former case, the chance is increased to keep the object in the field of view of the cameras. In the latter one, 3–D estimation becomes more reliable. Also, more details are provided, for example for recognizing the objects.

Beyond a rigorous mathematical framework we present real–time experiments demonstrating that we gain an improvement in 3–D trajectory estimation by up to 42% in comparison with tracking using a fixed focal length.

1. Introduction

This paper is not about a new tracking method. The main goal is to provide a framework for actively controlling the focal lengths of a camera pair while tracking a moving object in 3–D. Why are we interested in actively changing the focal length of cameras during tracking at all? At first, when considering surveillance tasks, like supervision of people in public buildings, the goal is not just to keep track of the moving person. It might also be necessary to identify people. And for identification (for example, face recognition) it is crucial to provide the algorithms with the highest resolution possible. Secondly, simple geometric considerations lead to the observation that 3–D estimation by means of observations in two image planes is improved in accuracy, if a larger focal length is used (assuming image noise being independent of the focal length). Summarizing, active focal length control during tracking can help to reduce uncertainty both in tracking and 3–D estimation, as well as in subsequent processing steps like recognition.

What are the problems when controlling the focal length during tracking? The main aspect, which we call *focal length dilemma* is the following: a larger focal length is usually preferred since more details of the moving objects are available in the image. At the same time, the risk is increased that the object is no longer completely visible or even totally out of the field of view of the camera. In other words, actively controlling the focal length makes it necessary to resolve this dilemma for each time step. The more knowledge about the 3–D position and trajectory of the moving object is available the larger the focal length can be set and as a consequence the more certain the estimation will be. On the other hand, the more uncertainty is remaining the smaller the focal length should be to avoid that the object is unexpectedly leaving the field of view of the camera.

To our knowledge no other work can be found on active focal length selection for improving accuracy in 3–D object tracking as stated above. In [7] focal length control is used to keep the size of the object constant during tracking, but without taking into account the uncertainty in the localization of the object. The work of [9] demonstrates how corner based tracking can be done while zooming using affine transfer. However, the focus is not on how to find the best focal length. Most related to our work are publications from

*This work was partially funded by the German Science Foundation (DFG) under grant SFB 603/TP B2. Only the authors are responsible for the content.

the area of active object recognition, where the best parameters of a camera (for example, position on a hemisphere around the object) is searched for, to reduce uncertainty in the recognition process [1, 3, 5, 11, 13, 14].

The theoretical foundations for our approach stem from a work on active object recognition and state estimation using information theoretic concepts [5]. The metric for sensor data selection, the mutual information between state and observation, results in the reduction of uncertainty in state estimation. In our contribution this metric is transferred to the case of focal length control for active object tracking. For that purpose, we embed the task of focal length control into the context of probabilistic state estimation. The notation of the Kalman filter usually applied in this area is slightly modified to handle time variant observation models (Sec. 2). This modification is required, since controlling the focal length inherently means changing the observation model over time. Applying the Kalman filter during state estimation yields a straight forward computation of the uncertainty in the state estimation. Three steps form the main impact of our work (Sec. 3):

1. the measure of uncertainty is derived from well founded information theoretic concepts. It is dependent of the focal length, so that we can influence the uncertainty by controling the focal length
2. the resulting metric can be computed a priori before any observation has been made from the moving object
3. thanks to the probabilistic modeling of the whole process we can reduce resolving the *focal length dilemma* to solving an optimization problem

The general mathematical formalism is applied to the case of tracking with two active cameras. Real–time experiments presented in Sec. 4 show that the active approach leads to an improvement in estimating the trajectory of a moving object in 3–D by up to 42%, compared with a strategy, where a fixed focal length is applied. Additionally, the approach shows the expected behavior of focal length control: for example, a large focal length is set, while the object stands still, and a small focal length, if the motion trajectory is not smooth and as a consequence the predicted position of the object at the next time step is very uncertain. This makes the approach valuable for further extensions, like combined tracking and recognition tasks, summarized in Sec. 5, where also further conclusions and an outlook to future work can be found.

2. Review: Kalman Filter for Changing Observation Models

In the following we consider a dynamic system whose state at time t is summarized by an n–dimensional state vector $\boldsymbol{x}_t$. The dynamic of the system is given by

$$\boldsymbol{x}_{t+1} = \boldsymbol{f}(\boldsymbol{x}_t, t) + \boldsymbol{w} \quad , \tag{1}$$

with $\boldsymbol{f}(\cdot,\cdot) \in \mathbb{R}^n$ being the state transition function and $\boldsymbol{w} \in \mathbb{R}^n$ being additive Gaussian noise with zero mean and covariance matrix $\boldsymbol{W}$. The observation $\boldsymbol{o}_t$ is given by the observation equation

$$\boldsymbol{o}_t = \boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{a}_t) + \boldsymbol{r} \quad , \tag{2}$$

which relates the state $\boldsymbol{x}_t$ to the observation $\boldsymbol{o}_t \in \mathbb{R}^m$. The function $\boldsymbol{h}(\cdot,\cdot) \in \mathbb{R}^m$ is called observation function and might incorporate the observations made by k different sensors. Again, an additive noise process $\boldsymbol{r} \in \mathbb{R}^m$ disturbs the ideal observation, with zero mean and covariance matrix $\boldsymbol{R}$.

The main difference to the standard description of a dynamic system and its occurrence in the world defined by the observation equation is the dependency of the observation function $\boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{a}_t)$ on the parameter $\boldsymbol{a}_t \in \mathbb{R}^l$. This vector summarizes all parameters that influence the sensor data acquisition process and as a consequence the observation $\boldsymbol{o}$ that is made by the sensors. One example for the parameter $\boldsymbol{a}_t$ might be $\boldsymbol{a}_t = (\alpha_t, \beta_t, f_t)^{\mathrm{T}}$, with the parameters α_t and β_t denoting the pan and tilt angles and the parameter f_t representing the motor controlled focal length of a multimedia camera at time step t.

For the time being let the parameter $\boldsymbol{a}_t$ be known and constant. State estimation of the dynamic system is performed by applying the standard Kalman filter approach. For simplicity we use the first order extended Kalman filter. The following notation is used:

- $\widehat{\boldsymbol{x}}_t^-$ is the predicted state estimate at time t without having made an observation.
- $\widehat{\boldsymbol{x}}_t^+$ is the state estimate at time t incorporating the observation made at time t.
- $\widehat{\boldsymbol{x}}_t$ is the state estimate at time t shortly before the state transition to time step $t+1$. The estimate is $\widehat{\boldsymbol{x}}_t = \widehat{\boldsymbol{x}}_t^+$ if the system has made an observation. The estimate equals $\widehat{\boldsymbol{x}}_t = \widehat{\boldsymbol{x}}_t^-$ if the system could not make an observation at time step t to update the state prediction.
- $\boldsymbol{P}_t^-$ is the covariance matrix for the predicted state $\widehat{\boldsymbol{x}}_t^-$, $\boldsymbol{P}_t^+$ the covariance matrix for the state estimate $\widehat{\boldsymbol{x}}_t^+$ after the observation. $\boldsymbol{P}_t$ is the covariance matrix of the state estimate $\widehat{\boldsymbol{x}}_t$ and will coincide with $\boldsymbol{P}_t^-$ or $\boldsymbol{P}_t^+$ depending on whether or not an observation could be made.

Using the defined quantities, the Kalman filter cycles through the following steps [2]:

1. State prediction

$$\widehat{\boldsymbol{x}}_t^- = \boldsymbol{f}(\widehat{\boldsymbol{x}}_{t-1}, t-1) \quad . \tag{3}$$

2. Covariance prediction

$$\boldsymbol{P}_t^- = \boldsymbol{f}_{\boldsymbol{x}} \boldsymbol{P}_{t-1} \boldsymbol{f}_{\boldsymbol{x}}^{\mathrm{T}} + \boldsymbol{W} \tag{4}$$

with $\boldsymbol{f}_{\boldsymbol{x}} = \left[\nabla_{\boldsymbol{x}} \boldsymbol{f}^{\mathrm{T}}(\boldsymbol{x}, t-1)\right]^{\mathrm{T}}_{\boldsymbol{x}=\widehat{\boldsymbol{x}}_{t-1}}$ being the Jacobian of the function $\boldsymbol{f}$ evaluated at the latest state estimate.

3. Filter gain computation

$$\boldsymbol{K} = \boldsymbol{P}_t^- \boldsymbol{h}_{\boldsymbol{x}}^{\mathrm{T}}(\boldsymbol{a}_t) \left(\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{a}_t) \boldsymbol{P}_t^- \boldsymbol{h}_{\boldsymbol{x}}^{\mathrm{T}}(\boldsymbol{a}_t) + \boldsymbol{R}\right)^{-1} \quad , \tag{5}$$

with $\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{a}_t) = \left[\nabla_{\boldsymbol{x}} \boldsymbol{h}^{\mathrm{T}}(\boldsymbol{x}, \boldsymbol{a}_t)\right]^{\mathrm{T}}_{\boldsymbol{x}=\widehat{\boldsymbol{x}}_t^-}$ being the Jacobian of the function $\boldsymbol{h}$ evaluated at the latest state estimate.

4. Update of state estimate (incorporation of observation $\boldsymbol{o}_t$)

$$\widehat{\boldsymbol{x}}_t^+ = \widehat{\boldsymbol{x}}_t^- + \boldsymbol{K}\left(\boldsymbol{o}_t - \boldsymbol{h}(\widehat{\boldsymbol{x}}_t^-, \boldsymbol{a}_t)\right) \quad . \tag{6}$$

5. State estimate covariance update

$$\boldsymbol{P}_t^+(\boldsymbol{a}_t) = \left(\boldsymbol{I} - \boldsymbol{K}\boldsymbol{h}_{\boldsymbol{x}}(\boldsymbol{a}_t)\right) \boldsymbol{P}_t^- \tag{7}$$

depending on the chosen parameter $\boldsymbol{a}_t$ that defines the observation function $\boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{a}_t)$.

The linearization by computing the Jacobian of the state transition function and the observation function introduces errors in the state estimate. There are several ways for reducing these errors (for example, using the second order extended Kalman filter [2]), or avoiding this linearization at all (for example, applying the iterated extended Kalman filter [2] or more modern approaches like particle filters [6, 10]). Here, we do not want to discuss the consequences and possible improvements for state estimation in general. Some remarks about combining our approach with other state estimators are given in the conclusion.

The linearization shown above allows us to model all distributions of the involved quantities $\boldsymbol{x}_t^-$, $\boldsymbol{o}_t$, and $\boldsymbol{x}_t^+$ as Gaussian distributions. Thus, we get the following distributions:

- A priori distribution over the state (and posterior, if no observation is made) $p(\boldsymbol{x}_t|\mathcal{O}_{t-1}, \mathcal{A}_{t-1}) \sim \mathcal{N}(\boldsymbol{x}_t^-, \boldsymbol{P}_t^-)$, with the two sets $\mathcal{A}_t = \{\boldsymbol{a}_t, \boldsymbol{a}_{t-1}, \ldots, \boldsymbol{a}_0\}$ and $\mathcal{O}_t = \{\boldsymbol{o}_t, \boldsymbol{o}_{t-1}, \ldots, \boldsymbol{o}_0\}$ denoting the history of actions and observations respectively.
- Likelihood function $p(\boldsymbol{o}_t|\boldsymbol{x}_t, \boldsymbol{a}_t) \sim \mathcal{N}(\boldsymbol{h}(\widehat{\boldsymbol{x}}_t^-, \boldsymbol{a}_t), \boldsymbol{R})$
- A posteriori distribution over the state (if an observation has been made) $p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t) \sim \mathcal{N}(\boldsymbol{x}_t^+, \boldsymbol{P}_t^+(\boldsymbol{a}_t))$

These three distributions are essential ingredients of our proposed optimality criterion, which is presented in the following.

3. Active Focal Length Control

In this section we develop a general optimality criterion for the selection of focal length parameters, which will result in largest reduction of uncertainty in the following estimation step. We like to stress that focal length control must be decided for *before* an observation is made. In other words, the criterion must not dependent on future observations. The proposed optimality criterion, given below in (9), shows exactly the postulated property. It is the conditional entropy of Gaussian distributed state and observation vectors. Here the reader can notice the benefits from the Kalman filter framework, summarized in the previous section.

3.1. Optimal Camera Parameters

The goal is to find an optimal camera parameter setting, i.e. the best parameter $\boldsymbol{a}$, that a priori reduces most the uncertainty in the state estimation with respect to the observation to be made in the following. In order to find the optimal camera parameters the important quantity to inspect is the posterior distribution. We want to improve state estimation by selecting the right sensor data. After we made an observation, we can exactly say how uncertain our state estimate is. Uncertainty is usually measured by the entropy $H(\boldsymbol{x}) = -\int p(\boldsymbol{x}) \log\left(p(\boldsymbol{x})\right) d\boldsymbol{x}$ of a random vector $\boldsymbol{x}$. Entropy can also be measured for a certain posterior distribution, for example for $p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t)$, resulting in

$$H(\boldsymbol{x}_t^+) = -\int p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t) \log\left(p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t)\right) d\boldsymbol{x}_t \quad .$$

This measure gives us *a posteriori* information about the uncertainty, if we took action $\boldsymbol{a}_t$ and observed $\boldsymbol{o}_t$. Deciding *a priori* about the expected uncertainty under a certain action $\boldsymbol{a}_t$ is of greater interest. The expected uncertainty can be calculated by

$$H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t) = \\ -\int p(\boldsymbol{o}_t|\boldsymbol{a}_t) \int p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t) \log\left(p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t)\right) d\boldsymbol{x}_t d\boldsymbol{o}_t.$$

The quantity $H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t)$ is called *conditional entropy* and depends in our case on the chosen parameter $\boldsymbol{a}_t$. Please note that the notation of the conditional entropy $H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t)$ is in accordance with information theory textbooks [4]. The quantity depends on the selected parameter vector $\boldsymbol{a}_t$, since this parameter will change the involved densities $p(\boldsymbol{o}_t|\boldsymbol{a}_t)$ and $p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t)$.

Now it is straight forward to ask the most important question for us: Which camera parameter yields the largest reduction of uncertainty? This question is answered by minimizing the conditional entropy for $\boldsymbol{a}_t$. In other words, the

best camera parameter $\boldsymbol{a}_t^*$ is given by

$$\boxed{\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmin}} H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t) \quad .} \tag{8}$$

Equation (8) defines the optimality criterion we have been seeking for in the case of arbitrary distributed state vectors. Unfortunately, in this general case the evaluation of (8) is not straightforward. Therefore, in the next section we consider a special class of distributions of the state vector, namely Gaussian distributed state vectors. This specialization allows us to combine the approach of camera parameter control with the Kalman filter framework and to compute the best action *a priori*.

3.2. Optimal Camera Parameters for Gaussian Distributed State Vectors

Resuming with the posterior distribution after the update of the state estimate in (6) we get the following conditional entropy:

$$H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t) = \int p(\boldsymbol{o}_t|\boldsymbol{a}_t) H(\boldsymbol{x}_t^+) d\boldsymbol{o}_t \quad .$$

As a consequence of the linearization in the extended Kalman filter, we know that the posterior distribution is Gaussian distributed. From information theory textbooks [4] we also know that the entropy of a Gaussian distributed random vector $\boldsymbol{x} \in \mathbb{R}^n$ with $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$H(\boldsymbol{x}) = \frac{n}{2} + \frac{1}{2} \log\left((2\pi)^n |\boldsymbol{\Sigma}|\right) \quad .$$

As a consequence, we get

$$H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t) = c + \int p(\boldsymbol{o}_t|\boldsymbol{a}_t) \frac{1}{2} \log\left(|\boldsymbol{P}_t^+(\boldsymbol{a}_t)|\right) d\boldsymbol{o}_t,$$

with c being a constant independent of $\boldsymbol{a}_t$. Now equation (8) becomes

$$\boxed{\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmin}} \int p(\boldsymbol{o}_t|\boldsymbol{a}_t) \log\left(|\boldsymbol{P}_t^+(\boldsymbol{a}_t)|\right) d\boldsymbol{o}_t \quad .} \tag{9}$$

From this equation we can conclude that we have to select the parameter $\boldsymbol{a}_t$ that minimizes the determinant of $\boldsymbol{P}_t^+(\boldsymbol{a}_t)$. Since $\boldsymbol{P}_t^+(\boldsymbol{a}_t)$ is independent of $\boldsymbol{o}_t$ for Gaussian distributed state vectors (compare (7)), the optimization can be done before the next observation is made. This was one of the main demands stated in the beginning of this section.

The optimization criterion in (9) is only valid if, for any chosen camera parameter $\boldsymbol{a}_t$, an observation can be made in any case. Obviously, this assumption is void, while arbitrarily changing the position and/or focal length of a camera. How to deal with this situation is considered in more detail in the next section.

3.3. Considering Visibility

Up to now we have assumed that at each time step an observation is made by the system to perform the state estimation update (6). Obviously, when changing the parameters of a sensor, depending on the state there is a certain a priori probability that no observation can be made that originates from the target. This has been denoted as focal length dilemma in the introduction. If no observation is made no update of the state estimate by means of (6) and (7) can be done. The resulting final state estimate for this time step is the predicted state estimate from the previous time step, with the corresponding predicted covariance matrix. The state prediction (cf. (3) and (4)) results in an increase of the covariance matrix since uncertainty is added, based on the dynamic of the system and the noise process disturbing the state transition process.

Tackling the focal length dilemma, the task of optimal sensor parameter selection can now be defined by finding a balance between the reduction of uncertainty in the state estimate and the risk of not making an observation and thus getting an increase of the uncertainty. Considering this trade–off in terms of the Kalman filter state estimation, the conditional entropy has to be rewritten as

$$\begin{aligned} &H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t) = \\ &= \underbrace{-\int_{\{v\}} p(\boldsymbol{o}_t|\boldsymbol{a}_t) H_v(\boldsymbol{x}_t^+) d\boldsymbol{o}_t}_{\text{target visible}} \underbrace{- \int_{\{\neg v\}} p(\boldsymbol{o}_t|\boldsymbol{a}_t) H_{\neg v}(\boldsymbol{x}_t^-) d\boldsymbol{o}_t}_{\text{target not visible}} . \end{aligned}$$

The first integral summarizes the entropy of the a posteriori probability for observations that can be made in the image (v). The probability of such observations weights the entropy $H_v(\boldsymbol{x}_t^+)$ of the a posteriori probability. The observations that cannot be measured in the image ($\neg v$) result in a Kalman filter cycle where no update of the state estimate is done and thus only a state prediction is possible. Then, the state prediction is treated as posterior. Again, the probability of such observations are used to weight the entropy $H_{\neg v}(\boldsymbol{x}_t^-)$ of the a posteriori probability. In the Kalman filter case, i.e. the estimation and propagation of Gaussian densities, the conditional entropy can further be simplified to a weighted sum

$$\begin{aligned} H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t) &= w_1(\boldsymbol{a}) \left(\frac{n}{2} + \frac{1}{2} \log\left((2\pi)^n |\boldsymbol{P}_t^+(\boldsymbol{a}_t)|\right) \right) \\ &+ w_2(\boldsymbol{a}) \left(\frac{n}{2} + \frac{1}{2} \log\left((2\pi)^n |\boldsymbol{P}_t^-|\right) \right) \quad . \end{aligned}$$

where the weights are given by

$$w_1(\boldsymbol{a}) = \int_{\{v\}} p(\boldsymbol{o}_t|\boldsymbol{a}) d\boldsymbol{o}_t \quad , \quad w_2(\boldsymbol{a}) = \int_{\{\neg v\}} p(\boldsymbol{o}_t|\boldsymbol{a}) d\boldsymbol{o}_t \quad . \tag{10}$$

For the minimization of $H_{\boldsymbol{a}_t}(\boldsymbol{x}_t|\boldsymbol{o}_t)$ the optimization problem is given by

$$\boxed{\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmin}} \left[w_1(\boldsymbol{a}) \log\left(|\boldsymbol{P}_t^+(\boldsymbol{a}_t)|\right) + w_2(\boldsymbol{a}) \log\left(|\boldsymbol{P}_t^-|\right)\right]} \tag{11}$$

One remark about the computation of the weights $w_1(\boldsymbol{a})$ and $w_2(\boldsymbol{a}) = 1 - w_1(\boldsymbol{a})$: since $p(\boldsymbol{o}_t|\boldsymbol{a})$ is Gaussian, in an implementation the weights can be computed based on the size of the sensor plane using the error function $erf(x) = \int_0^x e^{-0.5x^2}$.

Currently, we assume that the update of the state estimate is only done if all sensors observe the objects. For binocular object tracking conducted in the experiments such a 0–1 decision is sensible, since 3–D estimation based on 2D observations can only be done if the object is visible in both cameras. For the general case of k sensors the approach can be easily modified. A discussion of this topic is beyond the scope of this paper.

4. Real–time Experiments and Results

The following real–time experiments demonstrate the practicability and the benefits of our proposed method. It is shown that actively selecting the focal lengths increases the accuracy of the state estimation of a dynamic system.

We performed three different experiments:

1. fixating a static object while performing controlled movement of the binocular camera system, which is mounted on top of a mobile platform. Instead of the motion of the object the motion of the platform is estimated. The static object was located at a distance of approx. 2.7m. The platform was moving on a circle of diameter 0.6m.

2. tracking a toy train moving on a circular rail track. The distance to the object varied between 1.5m and 2.0m.

3. tracking a user controlled robot dog. The distance to the dog varied between 0.7m and 2.5m.

For the first two experiments we have inherently got ground truth data in 3–D, since we know the movement of the camera system (experiment 1) and the movement of the toy train (experiment 2). Thus quantitative evaluation of the tracking results (i.e. accuracy of 3–D estimation of the trajectory) is possible. For that reason, we performed two runs in each experiment, one with active camera control and the other with fixed focal lengths. In the third experiment, no ground truth data is available. With this experiment we show how the focal length is controlled in the case of unexpected movements, like sudden stops of the dog. Also, the movement of the object was not restricted to a periodic movement on a circular path.

4.1. Setup

For our experiments, we were using a calibrated binocular vision system (TRC Bisight/Unisight) equipped with two computer controlled zoom cameras, which is mounted on top of our mobile platform. In the following, tracking is done in a purely data driven manner without an explicit object model. Thus, at least two cameras are necessary to estimate the state (position, velocity, and acceleration) of the object in 3–D.

Both cameras look, slightly verging, into the same direction. The baseline between them is approx. 25 cm. We calibrated the cameras at 25 discrete zoom motor positions using Tsai's method [12] and stored the calibration parameters in a lookup table. The focal lengths range from approx. 17 mm to 38 mm. During tracking with zoom planning the focal length $\boldsymbol{a}$ is now not a continuous variable, but the number of one of the calibration data sets from the lookup table.

For the tracking itself, we used the region-based tracking algorithm proposed by Hager, et.al. [8], supplemented by a hierarchical approach to handle larger motions of the object between two successive frames. Given an initially defined reference template, the algorithm recursively estimates a transformation of the reference template to match the current appearance of the tracked object in the image. The appearance of the object might change due to motion of the object or due to changes in the imaging parameters. The advantage of this method is that it can directly handle scaling of the object's image region, which will appear while zooming. The reader should notice that any other tracking algorithm can be applied.

Finally, some remarks on the chosen state transition function (1) and observation equation (2) are given for the case of binoccular object tracking. The object is assumed to move with constant acceleration. Any other motion model is possible. For the state transition function (1) we use the linear model of a so called constant acceleration target [2]. The state vector $\boldsymbol{x}_t$ of such a dynamic system is given by

$$\boldsymbol{x}_t = (x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t, \ddot{x}_t, \ddot{y}_t, \ddot{z}_t)^{\mathrm{T}} \quad ,$$

with $(x_t, y_t, z_t)^{\mathrm{T}}$ being the position of the moving object in the world at time t. The non–linear observation equation

$$\boldsymbol{o}_t = (x_{L,t}, y_{L,t}, x_{R,t}, y_{R,t})^T = \boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{a}_t)$$

is defined by perspective projection of the world point $(x_t, y_t, z_t)^T$ to the image planes of both cameras. The parameter $\boldsymbol{a}_t = (f_L, f_R)^T$ summarizes the focal length of the left and right camera. The coordinates of the moving object in the image plane, returned by the tracking algorithm, are denoted as $x_{L,t}, y_{L,t}$ and $x_{R,t}, y_{R,t}$ for the left

and right camera, respectively. Since the observation equation is non–linear the Jacobian has to be used in the Kalman filter.

4.2. Experiment 1: Static Object — Moving Cameras

We conducted several real–time experiments that differ in the objects, in the backgrounds, and in the starting positions of the platform. As already mentioned, we performed two runs for each experiment, one with fixed focal lengths and one with active selection. In each case, the real–time binocular visual tracking is performed non–stop, even if the platform moves or the zoom motors are adjusted, but the state is estimated only when the platform stops between two successive moves. During the active tracking run and preceding each state estimation, the planning process starts and the zoom motors are adjusted according to (11). For the fixed case we chose the largest possible focal length that guarantees the visibility of the object for the whole experiment.

In Figure 1 images from one camera are shown taken during one of the experiments at approx. each twelfth planning step. The images give a visual impression of the planning results. We like to stress that the change in focal length is not driven by distance to the image border, that can easily be seen in image 6 and 12: although the object is close to the border of the image the estimation returns reliable velocity and acceleration values that indicate a movement of the object towards the image center.

Figure 1. Sample images from the left camera while tracking and zooming

The quantitative evaluation of the estimation error for the real–time experiments has been done by computing the mean squared error between the circular path and the estimated position. Averaged over all experiments, the mean squared error in the case of fixed focal lengths is $206.63\,\text{mm}$ (standard deviation: $76.08\,\text{mm}$) compared to an error of $154.93\,\text{mm}$ (standard deviation: $44.17\,\text{mm}$) while actively selecting the optimal focal lengths. This results in a reduction of the error by 25%. In Figure 2 the reconstructed movement path is shown for one of the experiments, comparing accuracy of 3–D estimation of the passive tracking approach with fixed focal lengths (Figure 2, left) with the active one (Figure 2, right).

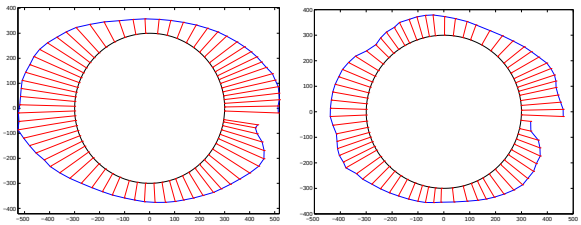

Figure 2. Visualization of real–time tracking estimation error. The inner circles represent the ground truth motion paths of the mobile platform. The outer curves show the estimated motion paths for fixed focal lengths (left) and for active zoom planning (right).

4.3. Experiment 2: Moving Toy Train — Static Camera

In the second experiment a moving toy train was tracked. For each experiment two runs have been executed: one with fixed focal lengths, one with actively changing the focal lengths according to our criterion. The focal lengths for the fixed case was selected in a way that the cameras could see the object during the whole movement on the circular track.

During the active tracking run every 5th image a new focal length is set. In Table 1 quantitative results for the estimation of the 3–D trajectory of the object are given. The results show that actively controlling the focal length during tracking reduces the estimation error in 3–D as well as the standard deviation of the error. In total, the reduction of the estimation error is up to 42%. In the average, the error is still reduced by 36%. We achieve a framerate of approx. 17 fps for both video streams and active focal length selection on an Athlon 1GHz processor. This shows that the approach can be utilized in real world applications. When we store the recorded images on disc during tracking, the framerate drops to 11–15 frames per second.

4.4. Experiment 3: Moving Robot Dog

The third experiment differs slightly from the previous two. In addition to the focal length control, the object has been fixated by the cameras during tracking using a PID

	passive		active	
	μ	σ	μ	σ
min	44.0591	21.4068	28.2702	15.1301
max	48.9534	26.5424	32.4285	18.7227
mean	47.1964	25.4800	30.1614	17.28083

Table 1. Estimation error of the 3–D trajectory of the moving object: passive vs. active approach. The best result (min), the worst result (max), and the result averaged over all experiments (mean) are given. Shown are the mean Euclidean estimation error in 3–D (μ) and the standard deviation (σ) per time step between the estimated movement and ground truth data (in mm). The distance to the moving object varied between 2.0m and 2.6m.

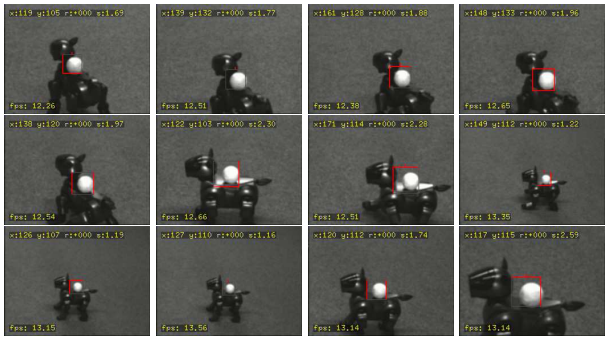

Figure 4. Sample images from the right camera while actively tracking the approaching dog. Every 50th frame is shown. The distance to the object is between 120cm (first images) and 70cm (last images).

controller. Fixation is done by setting the tilt axis of the stereo camera system and the vergence axes of the two cameras in a way that the object is kept in the center of the image.

The image sequence in Figure 3 demonstrates the expected behavior of an active focal length control. At first, while the object stands still the cameras zoom toward it (first image in Figure 3). Once the dogs starts moving backward (to the right in the image) the focal length is decreased in accordance with the remaining uncertainty in the state estimation (next four images). The dog stops again and the focal length thus is increased (images 6 and 7). Finally, the dog starts moving forward. As before the focal length is decreased (image 8). The process of focal length control can also be seen in Figure 5 (time steps between 5 and 30).

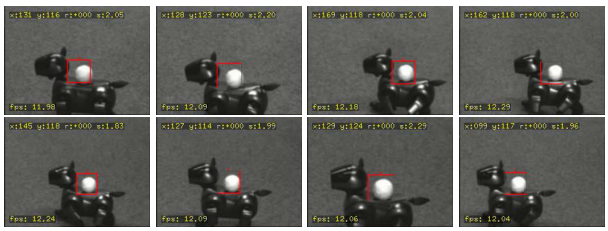

Figure 3. Sample images from the right camera while tracking and zooming. Every 50th frame of the recorded image sequence is shown. The distance is approx. 2.5m.[1]

The influence of the uncertainty in the 3–D estimation on the selected focal length can best be shown in Figure 4. The dog moves backwards and approaches the camera. Due to the constant movement, estimation is quite certain resulting in a large focal length. In image 6 the dog stops for a second and starts moving again to the right. This unexpected behavior causes the system to reduce the focal length quickly to the minimum value over the next 200 frames (approx. 7sec). Then the dog stops and the focal length can be increased for detailed inspection. The selected focal length can again be seen in Figure 5 for both cameras (time steps 50–85).

4.5. Summary

In three different experiments we have shown that object tracking will gain from the theoretically well founded approach for active focal length control. First, the 3–D estimation error is reduced in a significant way. Second, as a consequence of the criterion, the images are taken at the highest resolution possible. One important consequence of this property is that subsequent processing steps are provided always with most information about the moving target. Finally, the approach works in real time, which is one important demand for real world applications.

5. Conclusions and Future Work

In this paper we have presented an original approach on how to select the right focal length of two cameras in order to improve state estimation during object tracking. This problem has not been tackled before in the literature. The theoretically well founded criterion can be formulated for the general case of k sensors; it is not restricted to focal

[1]Movies can be downloaded from http://www5.informatik.uni-erlangen.de/MEDIA/denzler/ICCV03

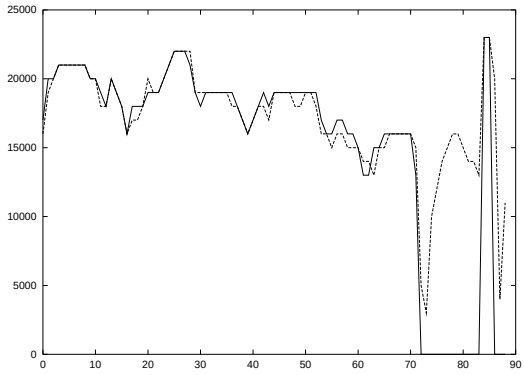

Figure 5. Plot of the focal length for the left and right camera during active tracking. The x–axis corresponds to the time steps (one time step corresponds to 200msec). The y–axis indicates motor positions for the focal length control (corresponds to a focal length between 17mm (value 0) and 38mm (value 25000).

length control only. For Gaussian distributed state vectors, a metric in closed form has been derived, which can be evaluated a priori and which can be optimized in real–time. We also showed how the whole approach fits into the Kalman filter framework and how to deal with the problem of visibility depending on the selected sensor parameters.

The approach has been verified and tested in real–time experiments for binocular object tracking. Active focal length control yields an improvement of up to 42% in the estimation error, compared to tracking, for which the focal length has been set constant. The whole approach runs at a framerate of approximately 17 fps on an Athlon 1GHz processor.

Besides the improvement in estimation, the approach is applicable beyond pure tracking tasks. Since the largest focal length with respect to the uncertainty is set, the images of the object have always the highest possible resolution in the current situation. Thus, continuative processing steps like object recognition will also gain from an active tracking strategy. In current work we investigate such combinations (tracking and recognition). Additionally, we work on integrating the uncertainty in the recognition process in the whole framework. The goal is to set the focal length not only based on the uncertainty in the estimation of the 3–D position of the moving object, but also based on the uncertainty in the recognition process.

Another point of interest is to apply the idea of information theoretic sensor data selection to the non–Gaussian case, which will become important if occlusions shall be handled by the tracking approach. This makes it necessary to use particle filters for state estimation and to optimize the criterion (11) for particle sets.

References

[1] T. Arbel and F.P. Ferrie. Viewpoint selection by navigation through entropy maps. In *Proceedings of the Seventh International Conference on Computer Vision*, pages 248–254, Kerkyra, Greece, 1999.

[2] Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press, Boston, San Diego, New York, 1988.

[3] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Appearance based active object recognition. *Image and Vision Computing*, (18):715–727, 2000.

[4] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley and Sons, New York, 1991.

[5] J. Denzler and C.M. Brown. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):145–157, 2002.

[6] A. Doucet, N. de Freitas, and N. Gordon, editors. *Sequential Monte Carlo Methods in Practice*. Springer, Berlin, 2001.

[7] J. Fayman, O. Sudarsky, and E. Rivlin. Zoom tracking and its applications. Technical Report CIS9717, Center for Intelligent Systems, Technion - Israel Institute of Technology, 1997.

[8] G.D. Hager and P.N. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

[9] E. Hayman, I. Reid, and D. Murray. Zooming while tracking using affine transfer. In *Proceedings of the 7th British Machine Vision Conference*, pages 395–404. BMVA Press, 1996.

[10] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998.

[11] B. Schiele and J.L. Crowley. Transinformation for active object recognition. In *Proceedings of the Sixth International Conference on Computer Vision*, pages 249–254, Bombay, India, 1998.

[12] R. Y. Tsai. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, Ra-3(3):323–344, August 1987.

[13] D. Wilkes. Active object recognition. Technical Report RBCV–TR–94–45, Department of Computer Science, University of Toronto, 1994.

[14] Y. Ye. Sensor planning for object search. Technical Report PhD Thesis, Department of Computer Science, University of Toronto, 1997.