On Optimal Camera Parameter Selection in Kalman Filter Based Object Tracking

Joachim Denzler, Matthias Zobel*, and Heinrich Niemann

Lehrstuhl für Mustererkennung, Universität Erlangen–Nürnberg
91058 Erlangen, Germany
{denzler, zobel, niemann}@informatik.uni-erlangen.de

Abstract In this paper we present an information theoretic framework that provides an optimality criterion for the selection of the best sensor data regarding state estimation of dynamic system. One relevant application in practice is tracking a moving object in 3–D using multiple sensors. Our approach extends previous and similar work in the area of active object recognition, i.e. state estimation of static systems. We derive a theoretically well founded metric based on the conditional entropy that is also close to intuition: select those camera parameters that result in sensor data containing most information for the following state estimation. In the case of state estimation with a non–linear Kalman filter we show how that metric can be evaluated in closed form.
The results of real–time experiments prove the benefits of our general approach in the case of active focal length adaption compared to fixed focal lengths. The main impact of the work consists in a uniform probabilistic description of sensor data selection, processing and fusion.

1 Introduction

In active vision it has been shown, also theoretically, that an active processing strategy that includes the image acquisition step can be superior to a passive one. The question of course is: how to figure out, which strategy is the best, i.e. how to actively control the image acquisition step in a theoretically optimal way? Answers to this question will have an enormous impact on broad areas in computer vision.

In recent work on active object recognition [2, 5, 11, 12] it has already been shown that sensible selection of viewpoints improves recognition rate, especially in settings, where multiple ambiguities between the objects exist. The work of [8] is another example from robotics, where such ideas have been implemented for self–localization tasks.

Besides the success in the area of active object recognition no comparable work is known for selecting the right sensor data during object tracking. The advantages are quite obvious:

– For a single camera setting, i.e. one camera is used to track the moving object, the trade–off between a large and a small focal length can be resolved. Depending on the position, velocity, acceleration of the object, and on the associated uncertainty in the estimation of these state values, a small focal length might be suited to ensure

* 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.

that the object is still in the image at the next time step. On the other hand, a large focal length can be advantageous for estimation of the class of an object in an high resolution image.

– For a setting with multiple cameras the possibility exists to focus with some cameras on dedicated areas in space, depending on the state estimation, while some other cameras keep on tracking using an overview image. Without mentioning the problems of sensor data fusion in such a setting and the weighting of the different sensors depending on the expected accuracy, the gain of dynamic focal length adjustment is expected to be relevant in a combined tracking and recognition scenario. Demonstrating this behavior is the focus of our ongoing research.

In this paper we make an important step toward active object tracking. The basis of our work is the information theoretic approach for iterative sensor data selection for state estimation in static systems presented in [4]. We extend this framework to have a metric for selecting that sensor data that yields most reduction of uncertainty in the state estimate of a dynamic system. We mainly differ to the work of [7, 10] in how to select the optimal focal length. In [7] zooming is used to keep the size of the object constant during tracking, but it is not taken into account the uncertainty in the localization. The work of [10] 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.

Our framework is completely embedded in probabilistic estimation theory. The main advantage is that it can be combined with any probabilistic state estimator. Also sensor data fusion, which is important in multiple camera settings, can be done at no extra costs. We show the realization for a two–camera setting in the experimental part of the paper later on. Another side effect of the information theoretic metric is its intuitive interpretation with respect to state estimation.

The paper is structured as follows. In the next section the main contribution of our work can be found. The general framework for selecting optimal observation models during state estimation of dynamic systems is applied to the case of binocular object tracking. Real–time experiments with a mobile platform in an office environment, presented in Section 3, give us definite, quantitative results: first, our optimality criterion causes the expected behavior, second, adapting the focal length during tracking is better than using a fixed setting, and third, the theoretical framework also works in practice. Some problems and future extension of our approach are discussed in Section 4.

2 Selection of Optimal Observation Models

2.1 Review: Kalman filter for Changing Observation Models

In the following we consider a dynamic system, whose state 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 a non–linear 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

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

which relates the state $\boldsymbol{x}_t$ to the observation $\boldsymbol{o}_t \in \mathbb{R}^m$. The non–linear function $\boldsymbol{h} \in \mathbb{R}^m$ 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, having zero mean and covariance matrix $\boldsymbol{R}$.

The main difference to the standard description of a dynamic system is the dependency of the observation function $\boldsymbol{h}(\boldsymbol{x}_t, \boldsymbol{a}_t, t)$ on the parameter $\boldsymbol{a}_t \in \mathbb{R}^l$. The vector $\boldsymbol{a}_t$ summarizes all parameters that influence the sensor data acquisition process. As a consequence the parameter also influences the observation $\boldsymbol{o}_t$, or in other words: we allow changing observation models. In the following the parameter $\boldsymbol{a}_t$ is referred to as action, to express that the observation model (i.e. the sensor data) is actively selected. One example for the parameter $\boldsymbol{a}_t$ might be $\boldsymbol{a}_t = (\alpha, \beta, f)^{\mathrm{T}}$, with α and β denoting the pan and tilt angles, and f the motor controlled focal length of a multimedia camera.

State estimation of the dynamic system in (1) and (2) can be performed by applying the standard non–linear Kalman filter approach. Space limitations restrain us from giving an introduction to the Kalman filter algorithm. The reader is referred to [1]. In the non–linear case given by (1) and (2) usually a linearization is done by computing the Jacobian of the state transition and the observation function. As a consequence of the linearization the distributions of the following random vectors are Gaussian distributed:

- A priori distribution over the state $\boldsymbol{x}_t$ (and posterior, if no observation is made)

$$p(\boldsymbol{x}_t|\mathcal{O}_{t-1}, \mathcal{A}_{t-1}) \sim \mathcal{N}(\boldsymbol{x}_t^-, \boldsymbol{P}_t^-) \quad ,$$

  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 $\boldsymbol{a}_t$ and observations $\boldsymbol{o}_t$ respectively. The quantities $\boldsymbol{x}_t^-$ and $\boldsymbol{P}_t^-$ are the predicted state and error covariance matrix, respectively [1].
- Likelihood function, i.e. the likelihood of the observation $\boldsymbol{o}_t$ given the state $\boldsymbol{x}_t$ and the action $\boldsymbol{a}_t$

$$p(\boldsymbol{o}_t|\boldsymbol{x}_t, \boldsymbol{a}_t) \sim \mathcal{N}(\boldsymbol{h}(\boldsymbol{x}_t^-, \boldsymbol{a}_t), \boldsymbol{R}) \quad .$$

- A posteriori distribution over the state space (if an observation has been made)

$$p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t) \sim \mathcal{N}(\boldsymbol{x}_t^+, \boldsymbol{P}_t^+(\boldsymbol{a}_t)) \quad . \tag{3}$$

  The vector $\boldsymbol{x}_t^+$ is the updated state estimate, the matrix $\boldsymbol{P}_t^+(\boldsymbol{a}_t)$ is the updated error covariance matrix. The matrix explicitly depends on the action $\boldsymbol{a}_t$ since the observation function in (2) depends on $\boldsymbol{a}_t$. In the case, that no observation has been made, the quantities $\boldsymbol{P}_t^+(\boldsymbol{a}_t)$ and $\boldsymbol{x}_t^+$ equal the corresponding predicted quantities $\boldsymbol{P}_t^-$ and $\boldsymbol{x}_t^-$.

These three distributions are essential ingredients of our proposed optimality criterion, which is presented in the following. Note, that all quantities (i.e. $\boldsymbol{x}_t^-, \boldsymbol{x}_t^+, \boldsymbol{P}_t^-, \boldsymbol{P}_t^+$) are updated in the Kalman filter framework over time [1]. As a consequence, this information is available at no extra cost for our approach.

2.2 Optimal Observation Models

The goal is to find an optimal observation model, i.e. the best action $\boldsymbol{a}_t$, that a priori most reduces the uncertainty in the state estimation with respect to the future observations. In order to find the optimal observation model the important quantity to inspect is

the posterior distribution. After an observation is made we can figure out, how uncertain our state estimate is. Uncertainty in a distribution of a random vector $\boldsymbol{x}$ can be measured by the entropy $H(\boldsymbol{x}) = -\int p(\boldsymbol{x}) \log(p(\boldsymbol{x}))\, d\boldsymbol{x}$. Entropy can also be calculated for a certain posterior distribution, for example for $p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t)$, resulting in

$$H(\boldsymbol{x}_t^+) = -\int p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t) \log\left(p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t)\right) d\boldsymbol{x}_t \quad . \tag{4}$$

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

$$H(\boldsymbol{x}_t|\boldsymbol{o}_t, \boldsymbol{a}_t) = -\int p(\boldsymbol{o}_t|\boldsymbol{a}_t) \int p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t) \log\left(p(\boldsymbol{x}_t|\mathcal{O}_t, \mathcal{A}_t)\right) d\boldsymbol{x}_t d\boldsymbol{o}_t \quad . \tag{5}$$

The quantity $H(\boldsymbol{x}_t|\boldsymbol{o}_t, \boldsymbol{a}_t)$ is called *conditional entropy* [3], and depends in our case on the chosen action $\boldsymbol{a}_t$. Having this quantity it is straight forward to ask the most important question for us: Which action yields the most reduction in uncertainty? The question is answered by minimizing the conditional entropy for $\boldsymbol{a}_t$, i.e. the best action $\boldsymbol{a}_t^*$ is given by

$$\boxed{\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmin}}\, H(\boldsymbol{x}_t|\boldsymbol{o}_t, \boldsymbol{a}_t) \quad .} \tag{6}$$

Equation (6) defines in the case of arbitrary distributed state vectors the optimality criterion we have been seeking for. Unfortunately, in the general case of arbitrary distributions the evaluation of (6) 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 selection of the best action with the Kalman filter framework. We will show, that this approach allows us to compute the best action *a priori*.

2.3 Optimal Observation Models for Gaussian Distributed State Vectors

We now continue with the posterior distribution in the Kalman filter framework. As a consequence of the linearization in the non–linear Kalman filter we know that the posterior distribution is Gaussian distributed (compare (3)). From information theory textbooks [3] 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 $H(\boldsymbol{x}) = \frac{n}{2} + \frac{1}{2}\log\left((2\pi)^n|\boldsymbol{\Sigma}|\right)$. Combining this knowledge we get for the conditional entropy $H(\boldsymbol{x}_t|\boldsymbol{o}_t, \boldsymbol{a}_t)$ of the distribution in (3)

$$H(\boldsymbol{x}_t|\boldsymbol{o}_t, \boldsymbol{a}_t) = \int p(\boldsymbol{o}_t|\boldsymbol{a}_t) \left(\frac{n}{2} + \frac{1}{2}\log\left((2\pi)^n|\boldsymbol{P}_t^+(\boldsymbol{a}_t)|\right)\right) d\boldsymbol{o}_t \quad . \tag{7}$$

Thus, neglecting the constant terms equation (6) becomes

$$\boxed{\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmin}} \int p(\boldsymbol{o}_t|\boldsymbol{a}_t) \log\left(|\boldsymbol{P}_t^+(\boldsymbol{a}_t)|\right) d\boldsymbol{o}_t \quad .} \tag{8}$$

From this equation we can conclude, that we have to select that action $\boldsymbol{a}_t$ that minimizes the determinant of $\boldsymbol{P}_t^+(\boldsymbol{a}_t)$. The key result is that $\boldsymbol{P}_t^+(\boldsymbol{a}_t)$ can be computed a priori, since the covariance matrix is independent of the observation $\boldsymbol{o}_t$.

The criterion in (8) is only valid in any case if for the chosen action $\boldsymbol{a}_t$ an observation from the system can be made. Obviously, in practice the selected action (i.e. camera parameter) will affect observability. How to deal with this situation is considered in more detail in the next section.

2.4 Considering Visibility

Up to now we have assumed that at each time step an observation is made to perform the state estimation update in the Kalman filter cycle. 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 dynamic system. An intuitive example is the selection of the focal length of a camera to track a moving object in the image. For certain focal lengths (depending on the 3–D position of the moving object) the object will no longer be visible in the image. As a consequence no observation is possible. How are time steps treated, for which no observations can be made, and what is the consequence for the state estimate? If no observation can be made, no update of the state estimate is possible. The resulting final state estimate for such a time step is the predicted state estimate $\boldsymbol{x}_t^-$, with the corresponding predicted covariance matrix $\boldsymbol{P}_t^-$. The implication on the state estimate is significant: during the prediction step in the Kalman filter algorithm the covariance matrix of the state estimate is increased; thus, uncertainty is added to the state estimate. The increase in uncertainty depends on the dynamic of the system and the noise process $\boldsymbol{w}$ disturbing the state transition process.

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

$$\boxed{H(\boldsymbol{x}_t|\boldsymbol{o}_t,\boldsymbol{a}_t) = \underbrace{\int\limits_{\{\text{visible}\}} p(\boldsymbol{o}_t|\boldsymbol{a})d\boldsymbol{o}_t}_{w_1(\boldsymbol{a})} H_v(\boldsymbol{x}_t^+) + \underbrace{\int\limits_{\{\text{invisible}\}} p(\boldsymbol{o}_t|\boldsymbol{a})d\boldsymbol{o}_t}_{w_2(\boldsymbol{a})} H_{\neg v}(\boldsymbol{x}_t^-),} \quad (9)$$

which is the weighted sum of $H_v(\boldsymbol{x}_t^+)$ and $H_{\neg v}(\boldsymbol{x}_t^-)$ where the weights are given by $w_1(\boldsymbol{a})$ and $w_2(\boldsymbol{a})$. The first integral in (9) summarizes the entropy of the a posteriori probability for observations that are generated by the system and that are *visible* in the image. The probability of such observations weight the entropy $H_v(\boldsymbol{x}_t^+)$ of the corresponding a posteriori probability (for simplifications in the notation, we use here $\boldsymbol{x}_t^+$ as synonym for the posterior). The observations that cannot be measured in the image (*invisible*) result in a Kalman filter cycle where no update of the state estimate is done and thus only a state prediction is possible. This state prediction is treated as a posteriori probability, without observation $\boldsymbol{o}_t$. Again, the probability of such observations are used to weight the entropy $H_{\neg v}(\boldsymbol{x}_t^-)$ of the a posteriori probability, when no observation has been made (again, we simplify notation by using $\boldsymbol{x}_t^-$ for the predicted state

distribution). Now the conditional entropy can be rewritten similar to (7). Thus, for the minimization of $H(\boldsymbol{x}_t|\boldsymbol{o}_t, \boldsymbol{a}_t)$ the optimization problem is given by

$$\boxed{\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmin}} \left[w_1(\boldsymbol{a}) \log\left(|\boldsymbol{P}_t^+(\boldsymbol{a}_t)|\right) + w_2(\boldsymbol{a}) \log\left(|\boldsymbol{P}_t^-|\right) \right] \quad .} \tag{10}$$

The minimization in (10) is done by Monte Carlo evaluation of the conditional entropy and discrete optimization. For more details on the derivation of the approach, on the optimization, and some special cases in practice the reader is referred to [6].

3 Real–time Experiments and Results

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

For our experiments we used a calibrated binocular vision system (TRC Bisight/ Unisight) equipped with two computer controlled zoom cameras that are mounted on top of our mobile platform. In the following, tracking is done in a pure 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.

In contrast to the usual setup for object tracking we did some kind of role reversal. Instead of a moving object with an unknown trajectory, we keep the object fixed at a certain position and track the object while moving the platform (with the mounted on cameras) on the floor in a defined manner. With this little trick, we obtain ground truth data from the odometry of the platform. It should be noted, that this information is not used for state estimation, but only for evaluation. For our experiments we decided to perform a circular motion with a radius of 300 mm. The object is located at a distance of about 2.7 m from the center of the circle in front of the platform. The optical axes of the two cameras are not parallel and lie not in the plane of the movement.

For the tracking itself, we used the region-based tracking algorithm proposed by [9], 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 for our demands is that it can directly handle scaling of the object's image region, as it will appear while zooming.

We conducted *three* real–time experiments that differ in the objects, in the backgrounds, and in the starting positions of the platform. We performed two runs for each experiment, one with fixed focal lengths and one with active selection. For the fixed case we chose the largest possible focal length that guarantees the visibility of the object for the whole run.

In Figure 1 images are shown from the left camera of the binocular camera system, taken during one of the experiments at approx. each twelfth planning step. The images give a visual impression of the planning results. As long as the uncertainty in the state estimate is high and the object is close to the border of the image the focal length is reduced (images 2 to 4, numbering line-by-line, starting top left). As soon as the

uncertainty in the state estimate (i.e. not only position, but also velocity and acceleration in 3–D) is low, the approach follows intuition and increases focal length, even in the case that the object is close to the border of the image (6 and 12). The reader should take into account, that not 2–D centering of the object in the image was our criterion for success of tracking. The goal was estimation of the movement path in 3–D by selecting the best focal length setting of the two cameras.

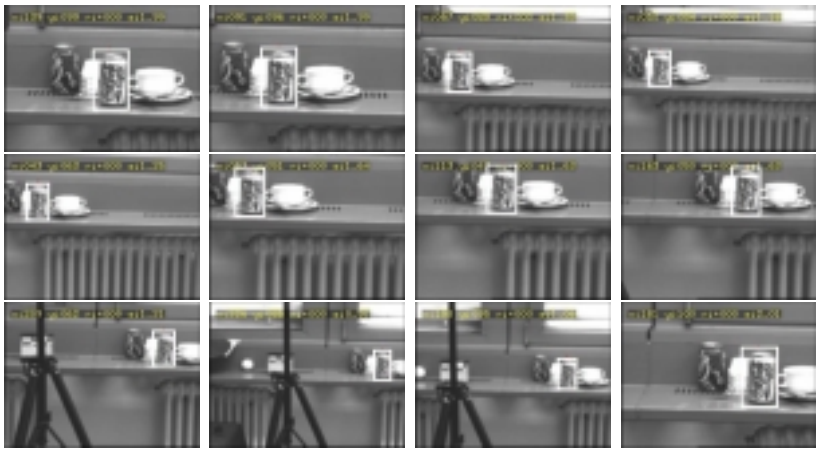

Figure 1. Sample images from the left camera while tracking the cola can and actively adjusting the focal length. Note, that it centering of the object was not the criterion during tacking!

The quantitative evaluation of the estimation error for the real–time experiments has been done by computing the Euclidean distance between the circular path and the estimated position. Averaged over all experiments, the mean distance in the case of fixed focal lengths is 206.63 mm (standard deviation: 76.08 mm) compared to an error of 154.93 mm (standard deviation: 44.17 mm) while actively selecting the optimal focal lengths. This results in a reduction in the error by 25%, as it has been similarly observed in simulations [6].

4 Conclusions and Future Work

In this paper we presented an original approach on how to select the right sensor data in order to improve state estimation of dynamic systems. For Gaussian distributed state vectors, a metric in closed form has been derived, that can be evaluated a priori. We 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. Although not discussed in this paper, the theoretically well founded criterion can be formulated for the general case of k sensors [6].

The main difference to previous work is that the selected focal length depends not only on the state estimate but also on the uncertainty of the state estimate and on the

reliability of the different sensors. Also, the special demands of the state estimator on the sensor data can be taken into account. This allows, for example, to solve the trade–off between large focal length for detailed inspection (for classification) and small focal length for tracking quickly moving objects. Experimental verification of this theoretical result are subject to future work.

The approach has been tested in real–time experiments for binocular object tracking. We tracked a static object while the cameras were moving. The estimated movement path of the camera is more accurate, when dynamically adapting the focal lengths. This result has been also verified in simulations, where the reduction in the estimation error was up to 43% [6].

Thanks to the Gaussian distributed state in the case of Kalman filter based tracking, the optimality criterion can be easily evaluated. However, for the general case of arbitrary distributed state vectors, the framework must be extended to allow the application of modern approaches like particle filters. In addition to that, we will verify the computational feasibility of our approach in applications, where frame–rate processing is necessary. One of the preliminaries to achieve frame–rate processing will be a smart and efficient way for the minimization in (6).

References

1. Y. Bar-Shalom and T.E. Fortmann. *Tracking and Data Association*. Academic Press, Boston, San Diego, New York, 1988.
2. H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Appearance based active object recognition. *Image and Vision Computing*, (18):715–727, 2000.
3. T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley and Sons, New York, 1991.
4. 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.
5. J. Denzler, C.M. Brown, and H. Niemann. Optimal Camera Parameter Selection for State Estimation with Applications in Object Recognition. In B. Radig and S. Florczyk, editors, *Pattern Recognition 2001*, pages 305–312, Berlin, September 2001. Springer.
6. J. Denzler and M. Zobel. On optimal observation models for kalman filter based tracking appoaches. Technical Report LME–TR–2001–03a, Lehrstuhl für Mustererkennung, Institut für Informatik, Universität Erlangen, 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. D. Fox, W. Burgard, and S. Thrun. Active markov localization for mobile robots. *Robotics and Autonomous Systems*, 25:195–207, 1998.
9. 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.
10. 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.
11. L. Paletta, M. Prantl, and A. Pinz. Learning temporal context in active object recognition using bayesian analysis. In *International Conference on Pattern Recognition*, volume 3, pages 695–699, Barcelona, 2000.
12. 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.