Optimal Camera Parameter Selection for State Estimation with Applications in Object Recognition

J. Denzler[1], C.M. Brown[2], and H. Niemann[1]

[1] Chair for Pattern Recognition, Department of Computer Science
Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen
denzler@informatik.uni-erlangen.de
http://www.mustererkennung.de
[2] Computer Science Department
University of Rochester, USA
brown@cs.rochester.edu
http://www.cs.rochester.edu

Abstract In this paper we introduce a formalism for optimal camera parameter selection for iterative state estimation. We consider a framework based on Shannon's information theory and select the camera parameters that maximize the mutual information, i.e. the information that the captured image conveys about the true state of the system. The technique explicitly takes into account the a priori probability governing the computation of the mutual information. Thus, a sequential decision process can be formed by treating the a posteriori probability at the current time step in the decision process as the a priori probability for the next time step. The convergence of the decision process can be proven.
We demonstrate the benefits of our approach using an active object recognition scenario. The results show that the sequential decision process outperforms a random strategy, both in the sense of recognition rate and number of views necessary to return a decision.

1 Introduction

State estimation from noisy image data is one of the key problems in computer vision. Besides the inherent difficulties with developing a state estimator that returns decent results in most situation, one important question is whether we can optimize state estimation by choosing the right sensor data as input. It is well known that the chosen sensor data has a big influence on the resulting state estimation. This general contiguity has been discussed in detail in dozens of papers in the area of active vision where the main goal was to select the right sensor data to solve a given problem.

In our paper we tackle the problem of optimal sensor data selection for state estimation by adjusting the camera parameters. The optimal camera parameters are found by minimizing the uncertainty and ambiguity in the state estimation process, given the sensor data. We will present a formal information theoretic metric for this informal characterization later on. We do not restrict the approach to acquiring sensor data once. The approach cycles through an action selection and sensor data acquisition stage where the sensor data decided for depends on the state estimation up to the current time step. One

important property of the proposed sequential decision process is that its convergence can be proven and that it is optimal in the sense of the reduction in uncertainty and ambiguity. We will demonstrate our approach in an active object recognition scenario.

The general problem of optimal sensor data acquisition has been discussed before. Examples can be found in the area of active robot localization [6] and active object recognition [1], where a similar metric has been used, but the sequential implementation is missing. The most related approach, not only from the application point of view, but also from a theoretical point of view is the work of [11]. The commonness, differences and improvements to this work are discussed later on. Similarities can also be found to the work of [2, 10], where a Bayesian approach [2] as well as an approach using reinforcement learning [10] has been presented for optimally selecting viewpoints in active object recognition. Our approach can be seen as a theoretically justifiable extension to this work. Interesting related work from the area of control theory are [9, 7].

The paper is structured as follows. In the next section we describe the problem in a formal manner and introduce the metric that is optimized during one step of the sequential decision process. In Section 3 we build up the sequential decision process and give a sketch of the convergence proof which can be found in detail in [4]. The active object recognition scenario is described in Section 4. The experimental results are summarized in Section 5. The paper concludes with a summary and an outlook to future work.

2 Formal Problem Statement

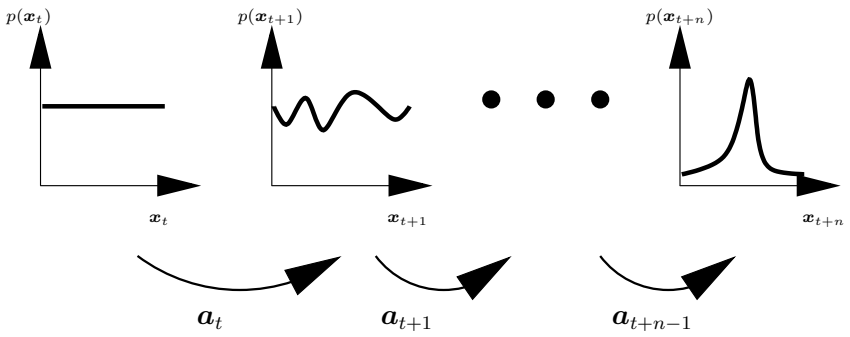

Figure 1. General principle: reduce uncertainty and ambiguity (variance and multiple modes) in the pdf of the state $\boldsymbol{x}_t$ by choosing appropriate information-acquisition actions $\boldsymbol{a}_t$.

The problem and the proposed solution are depicted in Figure 1. A sequence of probability density functions (pdf) $p(\boldsymbol{x}_t)$, $\boldsymbol{x}_t \in \mathcal{S}$ over the state space $\mathcal{S}$ is shown. The sequence starts with a uniform distribution, i.e. nothing is known about the state of the

system. Certain actions $\boldsymbol{a}_t$ are applied that select the sensor data at time step t. The following state estimation process results in a new probability distributions $p(\boldsymbol{x}_{t+1})$ over the state space. Finally, after n steps one should end up with a unimodal distribution with small variance and the mode close to the true state of the system. The problem now is twofold: first how to measure the success of a chosen action, i.e. how close the pdf is to a unimodal distribution with small variance. And second, how do we compute the action, that brings us closer to such a distribution.

The first question can be answered by using information theoretic concepts. In information theory the *entropy* of a pdf

$$H(\boldsymbol{x}_t) = -\int_{\boldsymbol{x}_t} p(\boldsymbol{x}_t)\log(p(\boldsymbol{x}_t))d\boldsymbol{x}_t$$

is defined which measures the amount of uncertainty in the outcome of a random experiment. The more unpredictable the outcome the larger the entropy is. It reaches its maximum for a uniform pdf and its minimum at zero for a delta function, i.e. for an unambiguous outcome.

The answer to the second question can also be found in information theory. Assume the following setting: the system is in state $\boldsymbol{x}_t$. The state itself cannot be observed but an observation $\boldsymbol{o}_t$ related with the state by a pdf $p(\boldsymbol{o}_t|\boldsymbol{x}_t, \boldsymbol{a}_t)$. The pdf is also conditioned on the action $\boldsymbol{a}_t$. In information theory the concept *mutual information* (MI) gives us a hint on which action $\boldsymbol{a}_t$ shall be chosen. The MI

$$I(\boldsymbol{x}_t; \boldsymbol{o}_t|\boldsymbol{a}_t) = \int_{\boldsymbol{x}_t}\int_{\boldsymbol{o}_t} p(\boldsymbol{x}_t)p(\boldsymbol{o}_t|\boldsymbol{x}_t, \boldsymbol{a}_t)\log\left(\frac{p(\boldsymbol{o}_t|\boldsymbol{x}_t, \boldsymbol{a}_t)}{p(\boldsymbol{o}_t|\boldsymbol{a}_t)}\right) d\boldsymbol{o}_t d\boldsymbol{x}_t \quad . \tag{1}$$

is the difference between the entropy $H(\boldsymbol{x}_t)$ and the conditional entropy $H(\boldsymbol{x}_t|\boldsymbol{o}_t, \boldsymbol{a}_t)$. It describes how much uncertainty is reduced in the mean about the true state $\boldsymbol{x}_t$ after the observation. Since we introduced the dependency on the action $\boldsymbol{a}_t$ we can influence the reduction in uncertainty by selecting that action $\boldsymbol{a}_t^*$ that maximizes the MI

$$\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmax}}\ I(\boldsymbol{x}_t; \boldsymbol{o}_t|\boldsymbol{a}_t) \quad . \tag{2}$$

All we need is the likelihood function $p(\boldsymbol{o}_t|\boldsymbol{x}_t, \boldsymbol{a}_t)$ and the a priori probability $p(\boldsymbol{x}_t)$.

In [11] a similar approach has been proposed in an active object recognition application, with the exception that the a priori information has been assumed to be uniform in any case. In the next section we extend this approach to a sequential decision process which convergence can be proven. The important difference is that we explicitly take into account the inherently changing prior. The prior changes, since new sensor data changes the information available about the true state.

3 Optimal Iterative Sensor Data Selection

From the previous section we know which action $\boldsymbol{a}_t$ to select to get the sensor data $\boldsymbol{o}_t$ that best reduces the uncertainty in the state estimation. From the definition of MI

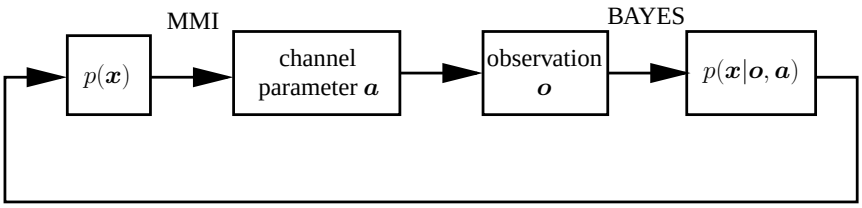

Figure 2. Sequential decision process of maximum mutual information (MMI) for camera parameter selection and Bayesian update of $p(\boldsymbol{x}|\boldsymbol{o},\boldsymbol{a})$ based on the observed feature $\boldsymbol{o}$.

it is obvious that the reduction will only be reached in the mean. As a consequence there might be observations under action $\boldsymbol{a}_t$ that result in an increase of the uncertainty. Another, more serious problem is, that there might be at more than one sensor data acquisition step necessary to resolve all ambiguity. An example is presented later on in the experimental section in the area of object recognition.

One way to deal with these problems is to form a sequential decision process and to take into account the information acquired so far, when selecting the next action. The sequential decision process consists of two step: the selection of the best action $\boldsymbol{a}_t$ based on the maximum of the mutual information (MMI) and the application of the Bayes rule to compute the a posterior probability when the observation has been made. The posterior is then fed back and used as prior for the next time step. This is justified by the fact that the posterior contains all information acquired to far, i.e. sensor data fusion is implicitly done during this step. In Figure 2 the whole sequential decision process is depicted.

By definition the iterative decision process is optimal since each step is optimal with respect to the prior of the state $\boldsymbol{x}_t$. Since the posterior is used as prior for the next time step we assure that the next action is selected considering the knowledge acquired so far. More important is the fact that this sequential decision process converges, i.e. the pdf $p(\boldsymbol{x})$ over the state space will converge towards a certain distribution. Only a sketch of the proof is given in the following.

It is known that any initial distribution over the Markov chain will converge to the unique stationary distribution, or to the minimum of all stationary distributions of the Markov chain. For a irreducible Markov chain at least one stationary distribution exists.

The key point of the convergence proof is that a irreducible Markov chain can be defined representing the sequential decision process [4]. Two corrolaries give us the proof of convergence. The first one is that the Kullback–Leibler distance between two distribution on a Markov chain will never increase over time. The second one is that the Kullback–Leibler distance between a distribution on a Markov chain and a stationary distribution on a Markov chain decreases over time. If there are more than one stationary distributions the convergence will be against the distribution with minimum Kullback–Leibler distance to all stationary distribution. Since each irreducible Markov chain has at least one stationary distribution we end up with a convergence toward a certain distribution over the Markov chain. This distribution is difficult to compute. But by this

result we know that the sequential decision process will converge. This convergence is important for practical considerations, i.e. when to stop the whole process.

In practice this convergence can also be observed. In many of our experiments in the area of active object recognition the distribution converges to the optimum distribution with respect to minimum entropy. Note, that it depends on the accuracy of the likelihood functions whether the resulting distribution will identify the right state. If the likelihood function, i.e. the relationship between state and observation, is erroneous, the sequential decision process cannot improve state estimation at all. On the one hand this might be seen as a drawback, since the state estimator is not optimized but only the sensor data provided for state estimation. On the other hand it is a big advantage, since any Bayesian state estimator at hand can be combined with the proposed sequential decision process. The more ambiguous the observations are the more the state estimation will be improved by our method.

Due to lack of space we have restricted us here to the description on the main principles. A more detailed discussion on the underlying information theoretic concepts as well as on the evaluation of the differential mutual information by Monte Carlo techniques can be found in [5]. There the reader will also find error bounds for the estimation of the mutual information.

4 Active Object Recognition Using Viewpoint Selection

To apply our proposed method we have chosen an object recognition scenario. We have selected a statistical Eigenspace approach which has been introduced in [2]. Here we apply it as the state estimator for classification.

The key idea is that the projection $\boldsymbol{c} = \boldsymbol{\Phi}_{\Omega_\kappa} \boldsymbol{f}$ of an image $\boldsymbol{f}$ into the Eigenspace of a class Ω_κ is assumed to be normally distributed, i.e. $p(\boldsymbol{c}|\boldsymbol{f}, \Omega_\kappa) \sim N(\boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa)$. Classification is then done not by computing the minimum distance in Eigenspace between a projected test image $\boldsymbol{f}$ and the manifold of a certain class [8] but by maximizing the a posteriori probability $\frac{1}{c} p(\boldsymbol{c}|\boldsymbol{f}, \Omega_\kappa) p(\Omega_\kappa)$. As a consequence the prior can be explicitly taken into account and one does not get only the best class hypotheses but also a statistical measure for the match. For viewpoint selection the likelihood functions $p(\boldsymbol{c}|\boldsymbol{f}, \boldsymbol{a}, \Omega_\kappa)$ for each viewpoint $\boldsymbol{a}$ have to be estimated during training. In our case a maximum likelihood estimation of the parameters of the Gaussian is performed. Due to lack of space only a coarse summary of the method can be given. More details are found in [2, 4, 5].

5 Experiments and Results

Five toy manikins form the data set (cf. Figure 3). There are only certain views from which the objects can be distinguished. The main differences in the objects are the small items that the manikins carry (band, lamp, quiver, gun, trumpet).

The experimental setup consists of a turntable and a robot arm with a camera mounted that can move around the turntable. The actions $\boldsymbol{a} = (\phi, \theta)^T$ define the position of the camera on the hemisphere around the object. The statistical eigenspace approach is used as classifier. The state $\boldsymbol{x}$ is the class of the object.

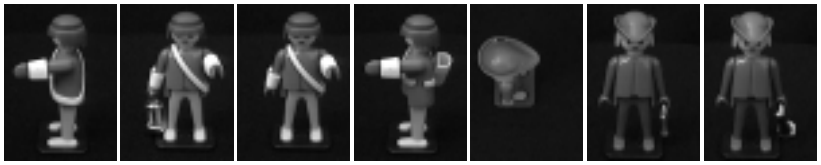

Figure 3. The first view is ambiguous with respect to the objects in image two and three. The second and third view allow for a distinction of objects two and three but not to distinguish object one from four (the objects with and without quiver on the back). Similar arguments hold for the two objects shown in the last three images.

We compared our viewpoint planning with a random strategy for viewpoint selection. Table 1 summarizes the results. The planning based on maximum mutual information outperforms a random strategy, in both recognition rate and number of views necessary for classification. In most cases the object can be recognized within three views at the latest. Also the maximum a posteriori probability after the decision for one class is larger in the mean for the viewpoint planning, indicating more confidence in the final decision (for example, object trumpet: 0.97 vs. 0.65). In contrast to other viewpoint selection approaches, for example based on reinforcement learning [3], we do not need to train the optimal sequence. All necessary information is already encoded in the likelihood functions, which are provided by the Bayesian classifier.

	planned viewpoint control			random viewpoint control		
object	rec. rate	mean no. views	mean max. a poster. prob.	rec. rate	mean no. views	mean max. a poster. prob.
band	86	1.13	0.98	77	4.28	0.95
lamp	97	1.14	0.98	93	4.94	0.96
quiver	99	1.05	0.99	95	3.09	0.97
gun	90	2.19	0.97	80	8.96	0.69
trumpet	99	2.29	0.97	70	8.89	0.65
average	94.2	1.56	0.97	83.0	6.03	0.84

Table 1. Results for viewpoint planning and random viewpoint control (100 trials per object): Recognition rate, mean number of views, and the mean of the maximum a posteriori probability for the right class after the decision.

In Figure 4 (left) the MI is shown at the beginning of the sequential decision process, i.e. the prior is assumed to be uniform. The x– and y–axis are the motorsteps for moving the turntable and the robot arm, to select positions of the camera on the hemisphere. The motorstep values correspond to a rotation between 0 and 360 degree for the turntable and -90 to 90 degree for the robot arm. The MI is computed by Monte Carlo simulation [5]. The maximum in this 2D function in the case of viewpoint selection defines the best

action (viewpoint) to be chosen. In Figure 4 (right) the corresponding view of the object is shown (for one of the objects as an example). This viewpoint is plausible, since the presence of the quiver as well as the lamp can be determined, so that three of the five objects can already be distinguished.

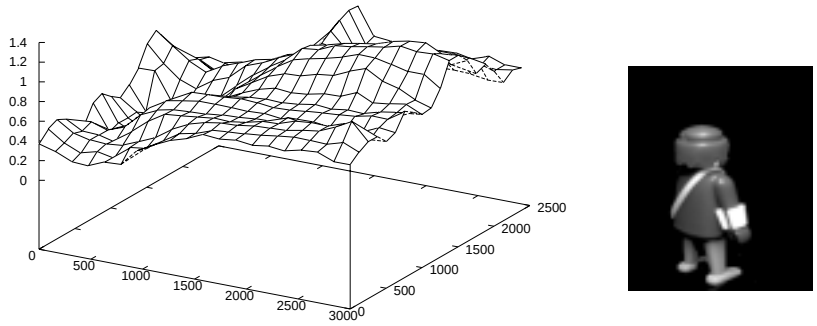

Figure 4. Left: MI in the viewpoint selection example assuming a uniform prior (computed by Monte Carlo evaluation). The x and y are the motorsteps for the turntable and robot arm, respectively. Along the z axis the MI is plotted. Right: best view $\boldsymbol{a}$ decided by the maximum in the MI ($\boldsymbol{a} = (2550, 1500)$). As example, object band is shown.

In general the computation time depends linearily on the number of actions and the number of classes. In practice, for viewpoint selection less than one second is needed on a Pentium II/300 for the computation of the best action using 1000 samples, 5 classes and a total of 3360 different actions (positions on the hemisphere).

6 Conclusion

We have presented a general framework for sensor data selection in state estimation. The approach has been applied to the optimal selection of camera parameters (viewpoint) in active object recognition. It is worth noting that the approach is not restricted to camera parameter selection but can be applied in any situation where the sensor acquisition process can be influenced. One examples is gaze control, where the pan/tilt/zoom parameters of a camera are changed [5]. Another example might be the adaptive change of illumination to enhance relevant features.

The approach presented in this paper is independent from the state estimator at hand. The only requirement is that the state estimator must provide likelihood functions for the observation given the state. The big advantage of this fact is, that any state

estimator can be improved by our method as long as the state estimator does not return systematically wrong results.

Compared to previously published work our approach forms a sequential decision process and its convergence can be proven. In contrast to reinforcement learning approaches [3] for active object reconition we do not need to train the optimal sequence. Thus, the typical tradeoff between exploitation and exploration in reinforcement learning does not exist for our framework. All relevant information necessary to decide for an optimal action is already encoded in the likelihood functions and the prior. The prior is computed step by step during the recognition process and the likelihood functions are assumed to be provided by the state estimator. Experiments showed that the framework works in an object recognition scenario with a state of the art classifier and outperforms a random strategy.

In our current work we extended this approach to state estimation of dynamic systems and we will modify the algorithms in a way that also continuous actions spaces can be handled.

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*, Kerkyra, Greece, 1999.
2. H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. Appearance based active object recognition. *Image and Vision Computing*, (18):715–727, 2000.
3. F. Deinzer, J. Denzler, and H. Niemann. Classifier Independent Viewpoint Selection for 3-D Object Recognition. In G. Sommer, N. Krüger, and Ch Perwass, editors, *Mustererkennung 2000*, pages 237–244, Berlin, September 2000. Springer.
4. J. Denzler and C. Brown. Optimal selection of camera parameters for state estimation of static systems: An information theoretic approach. Technical Report TR–732, Computer Science Department, University of Rochester, 2000.
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 (in press)*, 2001.
6. D. Fox, W. Burgard, and S. Thrun. Active markov localization for mobile robots. Technical report, Carnegie Mellon University, 1998.
7. J.M. Manyika and H.F. Durran-Whyte. On sensor management in decentralized data fusion. In *Proceedings of the Conference on Decision and Control*, pages 3506–3507, 1992.
8. H. Murase and S. Nayar. Visual Learning and Recognition of 3–D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, 1995.
9. C.A. Noonan and K.J. Orford. Entropy measures of multi–sensor fusion performance. In *Proceedings of the IEE Colloqium on Target Tracking and Data Fusion*, pages 15/1–15/5, 1996.
10. 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.
11. B. Schiele and J.L. Crowley. Transinformation for active object recognition. In *Proceedings of the Sixth International Conference on Computer Vision*, Bombay, India, 1998.