Information Theoretic Sensor Data Selection for Active Object Recognition and State Estimation

Joachim Denzler, *Member*, *IEEE*, and Christopher M. Brown

Abstract—We introduce a formalism for optimal sensor parameter selection for iterative state estimation in static systems. Our optimality criterion is the reduction of uncertainty in the state estimation *process*, rather than an estimator-specific metric (e.g., minimum mean squared estimate error). The claim is that state estimation becomes more reliable if the uncertainty and ambiguity in the estimation process can be reduced. We use Shannon's information theory to select information-gathering actions that maximize mutual information, thus optimizing the information that the data conveys about the true state of the system. The technique explicitly takes into account the a priori probabilities governing the computation of the mutual information. Thus, a sequential decision process can be formed by treating the a priori probability at a certain time step in the decision process as the a posteriori probability of the previous time step. We demonstrate the benefits of our approach in an object recognition application using an active camera for sequential gaze control and viewpoint selection. We describe experiments with discrete and continuous density representations that suggest the effectiveness of the approach.

Index Terms—Computer vision, active camera control, state estimation, information theory.

1 Introduction

The state, or state vector, of a system describes the relevant system parameters to be determined from observations by sensors. We use an information theoretic formulation to tackle the problem of optimal sensor data selection for state estimation. Many key problems in computer vision can be formulated as state estimation problems: For example, object classification (the state, i.e., the class of an object, is discrete and time independent), pose estimation (continuous and time independent state), and object tracking (the state is continuous and time variant).

Our ultimate goal is to provide a mechanism to select that sensor data that makes the state estimation minimally ambiguous and uncertain after interpreting the observations. Such a selection is very important since state estimation in computer vision is a process that always has to deal with uncertainties and ambiguities. Uncertainty arises from the noise in the sensor data, while ambiguity is based on inherent structure of the problem, e.g., objects identical in some views (Fig. 4).

In contrast to classical and modern approaches for state estimation [12], [4], our approach does not optimize a metric related to the state estimator, like its variance. Instead, we make use of the knowledge that is encoded in the state estimator as conditional probability densities. Uncertainty is improved not by changing the state estimator's knowledge, but by applying it in an optimal way in a sequential decision process. Optimality is defined in terms of reduction of uncertainty and ambiguity. A formal description of this kind of optimality is presented in Section 3.

The general principle and goal of our work is depicted in Fig. 1. A sequence of actions $\boldsymbol{a}_t$ is chosen in order to transform a prior distribution $p(\boldsymbol{x}_t)$ over the state space $\boldsymbol{x}_t \in \mathbb{R}^n$ ($p(\boldsymbol{x}_t)$ is uniform if no knowledge about the state is available) to a unimodal distribution with small variance whose mode uniquely identifies the right state. An action can be any controllable influence on image acquisition, data selection, or data processing. In a static system, the true state remains constant over time. In a dynamic system, the state changes over time following a dynamic model that itself is disturbed by noise. Although, our approach has in principle no restrictions that prevent it from being applied to dynamic problems (like zoom adjustments to track a moving object optimally) we focus here on static state estimation.

We demonstrate the benefits of the approach with an object recognition application, using active camera parameter selection. Here, there is a trade-off between detailed inspection and global overview that makes it difficult in general to choose an optimal focal length and viewing angle: A criterion is needed to balance this trade-off, using the current knowledge of the state. We have studied the adjustment of the focal length, the pan and tilt angles, and the camera's viewing position on a hemisphere around the object. The framework can be used for any other actions e.g., iris control or tuning of the focus of the camera.

- *J. Denzler is with the Lehrstuhl für Mustererkennung, Universität Erlangen-Nürnberg, Martensstr. 3, 91058 Erlangen, Germany. E-mail: denzler@informatik.uni-erlangen.de.*
- *C.M. Brown is with the Computer Science Department, University of Rochester, 160 Trustee Road, Rochester, NY 14627-0226. E-mail: brown@cs.rochester.edu.*

Manuscript received 26 Sept. 2000; revised 28 Feb. 2001; accepted 01 May 2001.
Recommended for acceptance by H.I. Christensen
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number 112911.

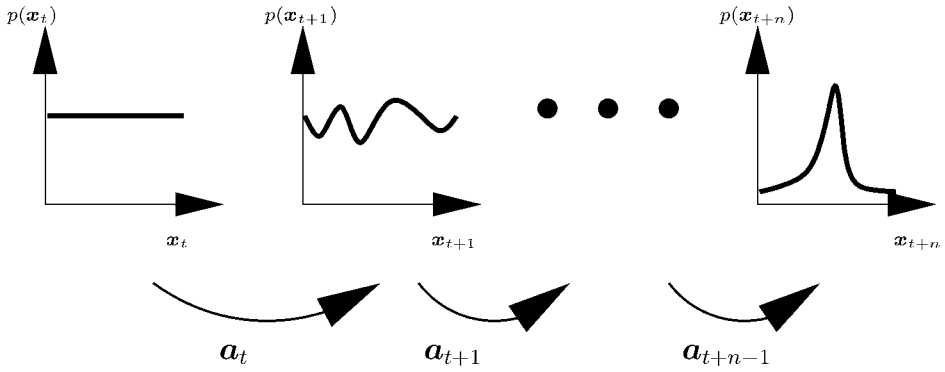


Fig. 1. General principle: reduce uncertainty and ambiguity (variance and multiple modes) in the pdf of the state x_t by choosing appropriate information-acquisition actions a_t.

The paper is structured as follows: We start with an overview of work applying information theoretic concepts in computer vision. Section 3 formally states the problem, which is tackled in Section 4 by a sequential decision process in the case of a time invariant system using an application of maximum mutual information (MMI) principle and Bayesian classification. Section 5 gives an example application for our framework, namely, active object recognition. The experimental evaluation is summarized in Section 6. The paper concludes with a discussion of results and a perspective on future work.

2 Related Work

Recently, the usefulness of information theoretic concepts in computer vision has been recognized, with application in tasks like image registration [23], viewpoint selection in object recognition [19], [2], and feature extraction [8].

Our motivation and starting point is the approach to active object recognition described in [19]. That work remains the closest to our approach. In [19], an active object recognition scheme uses transinformation (mutual information, MI) to place receptive fields optimally over the object of interest. We also use MI, but the contribution of our work is in the extension to a sequential decision process whose convergence can be proven. Technically, the main feature of our extension is that we explicitly take the change in the prior distrbution into account, while [19] assumes an unchanging, uniform prior. However, they perform both classification and localization of objects in 3D.

Another information-theoretic approach with remarkable results in viewpoint selection [2], [3], [18], [17] is related to, but significantly different from, ours. In contrast to our metric, described in detail in Section 4, this prior work uses the "average loss in entropy"

$$E[H(\boldsymbol{x}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_n) - H(\boldsymbol{x}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_n,\boldsymbol{o}_{n+1},\boldsymbol{a})] \quad (1)$$

as the metric to optimize. This quantity is closely related to the expected value of the MI: This is because the term

$$H(\boldsymbol{x}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_n) - H(\boldsymbol{x}|\boldsymbol{o}_1,\ldots,\boldsymbol{o}_n,\boldsymbol{o}_{n+1},\boldsymbol{a}) \quad (2)$$

is equal to the MI as long as the second term is the conditional entropy [2, (11)]. However, [18], [17] defines

$$H(\boldsymbol{x}|\boldsymbol{o}) = \sum p(\boldsymbol{x}|\boldsymbol{o}) \log p(\boldsymbol{x}|\boldsymbol{o}) \quad (3)$$

i.e., as the entropy of the posterior. This quantity differs from the conditional entropy, thus, leading to a quantity that is not exactly the expected value of the MI.

These technical differences in the information-theoretic metric (our MI versus non-MI or MI-average metrics) result from our commitment to sequential decision making, for which MI is the theoretically justifiable metric. The MI approach yields a natural solution to the problems that [3], [2] report with views from similar viewpoints (verified in our experiments). In [3], this problem is solved by the heuristic of masking out already visited viewpoints. Finally, since we can prove the convergence of our proposed sequential decision process, it is unnecessary to define an experimentally adjusted empirical threshold for stopping the view planning, as it is done in [17]. We believe that there is no need to compute the average MI and we do not believe that the use of the the difference between the entropy of the prior and posterior probability holds in the general case.

Interestingly, this latter metric is a suitable reward function in the reinforcement learning [18], [17], in which Q-learning or other approximate techniques are used to solve Bellman's equations [20]. This reward is possible since one observes the reward *after* the performed actions. Based on this information, an action selection mechanism is learned during trial and error steps [20]. In our work, we want to know the best action *before* we perform it and the necessary information for this is collected during a training step where we estimate the likelihood functions $p(\boldsymbol{o}|\boldsymbol{x},\boldsymbol{a})$ for a given action. The MI tells us which action $\boldsymbol{a}$ is best, based on the prior and the likelihood functions (details in Section 4). One drawback of the reinforcement learning approach is that it needs the learning step, in which the action selection mechanism is trained by a trial and error method. If one already has the conditional distribution in (6), no training step is necessary for the action selection. All necessary information is already encoded in the conditional distributions and can be applied directly. However, if no statistical classifier is applied, reinforcement learning provides a suitable mechanism (if not the best one) to find the best action. But then, sensor data fusion, as is done

easily by applying Bayes rule in [18], [17], remains a serious problem.

In the area of image registration, the work of [23] is a good example of the rigorous application of information theoretic concepts in computer vision. The alignment of two images that do not necessarily come from the same modality is done by maximizing the MI. This theoretically complicated and practically expensive step is elegantly performed with the stochastic optimization algorithm EMMA. The underlying pdf's are represented by Parzen window densities. The authors also show applications in the area of object tracking and photometric stereo. These techniques have parallels in principal component analysis and function learning [24].

In [8], an information theoretic approach for feature extraction motivated by Fano's inequality for the error rate in classification is presented. This work also represents the continuous pdf's by Parzen window densities. It can be seen as a practical realization of a feature selection scheme based on the MI; in fact, it can also be found in textbooks on pattern recognition [15]. Related to our work, the approach in [8] covers one step of our sequential decision process.

Information theoretic concepts have been investigated recently for active vision and action selection. Examples are active localization of robots [9], active view point selection for object recognition [1], and sensor planning for active object search [26]. In an interesting, but philosophically different approach, [25], optimal ("special") views are defined as those that allow the best feature matching even in the case of occlusion. In addition to the general investigation of a geometrical definition of a special view that is dependent of the classifier chosen, one of the main results is a strategy for how to determine these special views using image data only.

In the control literature, two interesting contributions use the entropy for optimal state estimation under the special aspect of sensor fusion. In [16], the main concept is to use entropy for sensor fusion in Kalman-filter-like settings, i.e., Gaussian noise and linear dynamic systems. The relationship between the Fisher information matrix and the Cramer-Rao lower bound on the error in state estimation is applied to sensor fusion. Simulated 3D tracking leads to "information maps," which show the best positions for two sensors capable of measuring only the direction of a moving object. Closely related work was first presented in [13], where the main focus is on sensor management in data fusion. Again, the entropy is taken as a measure of information. Although, neither approach can be directly applied to our problem, they give very useful hints on how to extend and formulate the optimal sensor data selection in state estimation of linear, dynamic systems with Gaussian noise.

Although our work is in the mainstream of information-theoretic research, we believe that there are important differences with other work with which we are familiar. Our main contribution consists of the description of a complete framework for sensor data selection based on Shannon's information theory (Sections 3 and 4). This framework directly points to the metric to be optimized, the MI. Interpreting the posterior as prior for the next time step's sensor, implicitly performs data fusion and a sequential decision process emerges whose convergence can be formally proven. Further, the entire approach can be extended to continuous distributions through the use of differential MI (Section 4.3).

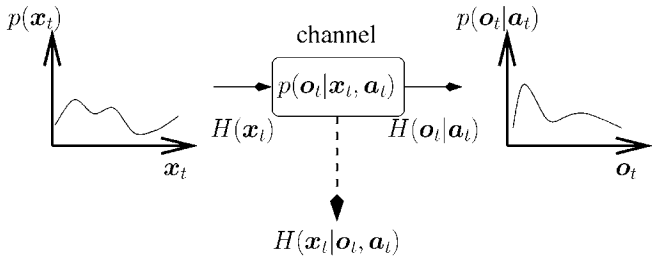


Fig. 2. Input and output relation in the channel model and some of the important entropies, H, describing the information content. The state estimator that estimates the state $\boldsymbol{x}_t$ based on the observation $\boldsymbol{o}_t$ is missing in this figure.

3 Formal Problem Statement

Most problems in computer vision, especially dynamic problems, cycle (either explicitly or implicitly) through a state estimation and action selection stage. Based on the image data $\boldsymbol{o}_t$ or some other acquired sensor information at time step t the unobservable true state $\boldsymbol{x}_t$ of the system (static or time varying) is approximated by a state estimate $\widehat{\boldsymbol{x}}_t$. This estimated state is the basis for selecting a certain action $\boldsymbol{a}_t$, which is performed in order to reach a predefined goal. For a static system, a goal might be to improve state estimation by using additional sensor data, which ideally should be selected optimally. The goal in a dynamic system might be to reduce the error between the estimated and true state over time or to make the pdf of the state as much like a delta function as possible.

We use a probabilistic framework: Sensor data is not noiseless or ideal, nor can the effectiveness of actions be known in advance. In a probabilistic framework, this uncertainty can be modeled by adding a stochastic noise component to the parameters that must be estimated. Noise estimation can be done during training, or by making assumptions that are checked and adjusted during system operation.

In object tracking, the (time-varying) state of the system could be the position, velocity, and acceleration of the object in 3D and an action could be the selection of pan and tilt movements needed to keep the moving object in the image. In object recognition, the (static) state of the system is the class of the object and the actions might be camera movements to reach optimal disambiguating viewpoints [2], [19].

Fig. 2 gives the main elements of our approach. It shows the transmission of a state $\boldsymbol{x}_t$ over a channel. At the other end of the channel, an observation $\boldsymbol{o}_t$ is made. The system gets as input an a priori distribution over the state space $p(\boldsymbol{x}_t|\boldsymbol{o}_{t-1},\ldots,\boldsymbol{o}_0)$ that describes the belief of being in a certain state $\boldsymbol{x}_t$ at time t given that the previous sensor readings have been $\boldsymbol{o}_{t-1},\boldsymbol{o}_{t-2},\ldots,\boldsymbol{o}_0$. In Fig. 2, we have left out the dependency on the past observations in $p(\boldsymbol{x}_t|\boldsymbol{o}_{t-1},\ldots,\boldsymbol{o}_0)$ for

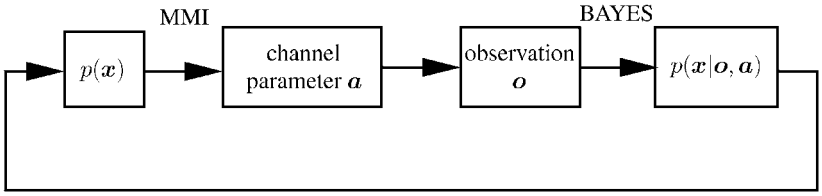


Fig. 3. Sequential decision process of maximum mutual information (MMI) for camera parameter selection and Bayesian update of $p(\boldsymbol{x}|o,\boldsymbol{a})$ based on the observed feature o.

clarity. For a static system, the distribution is equal to $p(\boldsymbol{x}_{t-1}|\boldsymbol{o}_{t-1},\ldots,\boldsymbol{o}_0)$. In a dynamic system, $p(\boldsymbol{x}_t|\boldsymbol{o}_{t-1},\ldots,\boldsymbol{o}_0)$ is calculated by

$$p(\boldsymbol{x}_t|\boldsymbol{o}_{t-1},\ldots,\boldsymbol{o}_0) = \int_{\boldsymbol{x}_{t-1}} p(\boldsymbol{x}_{t-1}|\boldsymbol{o}_{t-1},\ldots,\boldsymbol{o}_0)p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})d\boldsymbol{x}_{t-1} \tag{4}$$

using a model $p(\boldsymbol{x}_t|\boldsymbol{x}_{t-1})$ of the dynamics of the system. With that pdf, an entropy

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

is associated (definitions of relevant information-theoretic terms can be found in [7], [5]). The entropy measures the amount of uncertainty in a random experiment using the pdf $p(\boldsymbol{x}_t)$. The entropy is zero if the outcome of the experiment is unambiguous; it reaches its maximum if all outcomes of the experiment are equally likely.

The true state $\boldsymbol{x}_t$ cannot be observed. Following the information theoretic formulation, the state is sent through the channel, whose parameter is summarized by $\boldsymbol{a}$. The transmission over the channel can be interpreted as the image formation process. On the other end of the channel an observation $\boldsymbol{o}_t$ is received. The observation is related to the state by the likelihood function $p(\boldsymbol{o}_t|\boldsymbol{x}_t,\boldsymbol{a}_t)$, which is proportional to the probability that an observation $\boldsymbol{o}_t$ is made if the state $\boldsymbol{x}_t$ is sent through the channel. The likelihood function also serves as a model of the noise component in the channel; for example, $p(\boldsymbol{o}_t|\boldsymbol{x}_t,\boldsymbol{a}_t)$ might be a Gaussian distribution with mean value $\boldsymbol{x}_t$ and variance depending on the chosen action $\boldsymbol{a}_t$ or on both the state $\boldsymbol{x}_t$ and the action $\boldsymbol{a}_t$. The pdf $p(\boldsymbol{o}_t|\boldsymbol{a}_t)$ of the observation is defined as

$$p(\boldsymbol{o}_t|\boldsymbol{a}_t) = \int_{\boldsymbol{x}_t} p(\boldsymbol{o}_t|\boldsymbol{x}_t,\boldsymbol{a}_t)p(\boldsymbol{x}_t)d\boldsymbol{x}_t.$$

Again, an entropy $H(\boldsymbol{o}_t|\boldsymbol{a}_t)$ can be associated with the distribution $p(\boldsymbol{o}_t|\boldsymbol{a}_t)$. The important quantity in this formalism is the chosen action $\boldsymbol{a}_t$. Since the likelihood function $p(\boldsymbol{o}_t|\boldsymbol{x}_t,\boldsymbol{a}_t)$ is conditioned on this action, the action itself influences the properties of the channel. Still, the goal is to estimate the true state $\boldsymbol{x}_t$, given the observation $\boldsymbol{o}_t$. In information theory, *mutual information (MI)* (or *transinformation*) defines how much uncertainty is reduced in $\boldsymbol{x}_t$ if the observation $\boldsymbol{o}_t$ is made. Since the information flow through the channel depends on the parameter $\boldsymbol{a}_t$, we need to define conditional MI as

$$I(\boldsymbol{x}_t;\boldsymbol{o}_t|\boldsymbol{a}_t) = H(\boldsymbol{x}_t) - H(\boldsymbol{x}_t|\boldsymbol{o}_t,\boldsymbol{a}_t). \tag{5}$$

Some properties of the MI are discussed in [7]. Using the above notation for the conditional probabilities and the definition of the entropies $H(\boldsymbol{x}_t)$ and $H(\boldsymbol{x}_t|\boldsymbol{o}_t,\boldsymbol{a}_t)$, the MI becomes

$$I(\boldsymbol{x}_t;\boldsymbol{o}_t|\boldsymbol{a}_t) = \int_{\boldsymbol{x}_t}\int_{\boldsymbol{o}_t} p(\boldsymbol{x}_t)p(\boldsymbol{o}_t|\boldsymbol{x}_t,\boldsymbol{a}_t)\log\left(\frac{p(\boldsymbol{o}_t|\boldsymbol{x}_t,\boldsymbol{a}_t)}{p(\boldsymbol{o}_t|\boldsymbol{a}_t)}\right)d\boldsymbol{o}_t d\boldsymbol{x}_t. \tag{6}$$

Since we are interested in reducing the uncertainty, if the state is sent through the channel and an observation is made on the other end of the channel, we have to maximize the MI. The MI is a function of the parameter $\boldsymbol{a}_t$ and, thus, the optimal action $\boldsymbol{a}_t^*$, given an a priori distribution $p(\boldsymbol{x}_t)$ and a model for the channel noise $p(\boldsymbol{o}_t|\boldsymbol{x}_t,\boldsymbol{a}_t)$, is

$$\boldsymbol{a}_t^* = \underset{\boldsymbol{a}_t}{\operatorname{argmax}}\ I(\boldsymbol{x}_t;\boldsymbol{o}_t|\boldsymbol{a}_t). \tag{7}$$

4 Sequential Decision Making

Our MI framework leads naturally to an iterative algorithm for state estimation. The general formulation of a provably-convergent sequential decision process for optimal sensor data selection is the main contribution of our paper. The relation to previously published work is given in Section 2. Continuous random variables and Monte Carlo techniques are addressed in Section 4.2 and Section 4.3.

The use of the MI allows a recursive evaluation and judgment of the next viewpoint and thus forms a sequential decision process, as shown in Fig. 3.

At the beginning of the sequential decision process (say, at time $t = 0$), the a priori probability over the state space $p(\boldsymbol{x}_0)$ is initialized (uniformly, unless reliable, nonuniform priors are known). The first camera parameter $\boldsymbol{a}_0$ is selected by maximizing the MI (7). The resulting image (using camera parameter $\boldsymbol{a}_0$) or some information extracted from this image serves as observation $\boldsymbol{o}_0$. Bayes rule returns the a posteriori probability $p(\boldsymbol{x}_0|\boldsymbol{o}_0,\boldsymbol{a}_0) = \frac{p(\boldsymbol{o}_0|\boldsymbol{x},\boldsymbol{a}_0)p(\boldsymbol{x}_0)}{p(\boldsymbol{o}_0|\boldsymbol{a}_0)}$, justified by the fact that the a priori probability does not depend on the chosen camera parameters.

The computed a posteriori probabilities can be interpreted as new a priori probabilities for the next view-planning step, i.e., $p(\boldsymbol{x}_1) = p(\boldsymbol{x}_0|\boldsymbol{o}_0,\boldsymbol{a}_0)$. As a consequence, the MI in (6) will change after the first update of the state estimate. In general, after the nth view-planning step, one gets as prior probability of time step $n+1$,

$$p(\boldsymbol{x}_{n+1}) = \frac{p(\boldsymbol{o}_n|\boldsymbol{x}_n, \boldsymbol{a}_n)p(\boldsymbol{x}_n|\boldsymbol{o}_{n-1}, \ldots, \boldsymbol{o}_0)}{p(\boldsymbol{o}_n|\boldsymbol{a}_n)}, \tag{8}$$

and

$$\boldsymbol{a}_n = \underset{\boldsymbol{a}}{\operatorname{argmax}}\, I(\boldsymbol{x}_n; \boldsymbol{o}|\boldsymbol{a}). \tag{9}$$

Here, the plausible assumption is made that the distribution of the features of view n depends only on the class and the chosen view, but not on the past views and that the properties of the channel, i.e., the likelihood function, does not change. Equations (8) and (9) define the process of recursive viewpoint selection.

Formulating sequential decisions as a Markov decision process (MDP) suggests dynamic programming, the technique that is the basis of most algorithms for configuring MDPs from examples (see, for example, the textbook on reinforcement learning by Sutton and Barto [20]). Recently, the partially observable (POMDP) case [11] has been treated, but still by either applying dynamic programming or directly solving the Bellman equations. Our method avoids time and memory-intensive dynamic programming. However, in our approach, the estimation of the necessary statistical information (6) is not a trivial task. Ideally, this estimation could be unnecessary if such knowledge is provided by the state estimator.

4.1 Convergence and Optimality of the Sequential Decision Making

The experiments in Section 6, show that the sequential decision making process converges in practice. Actually, this convergence can also be formally proven [7]. One consequence of the proof is that, under certain difficult-to-verify conditions, the sequential decision process is also guaranteed to identify the true state. Under general conditions, proving this remains an unsolved problem.

What can be proven is the optimality in the sense of reduction in uncertainty. Since the MI for a fixed a priori probability depends only on the conditional entropy, i.e., the mean value of the entropy of the a posteriori probability averaged over all possible observations, maximizing the MI means minimizing the conditional entropy (compare (5)). This follows directly from the definition of MI. One cannot assure that, for every single step in the sequential decision process, the uncertainty is reduced. The change in uncertainty depends on the current observation. On average i.e., in the long run, by definition of the MI, the uncertainty will be reduced, which was one of our main goals defined in Section 1.

4.2 Differential Entropy and Mutual Information

A discrete representation of the pdf's simplifies the evaluation of the MI. We now extend the sequential decision process to use MI evaluated from a continuous pdf.

The differential entropy $h(x)$ of a continuous random variable x with density $p(x)$ is defined as [5]

$$h(x) = -\int p(x)\log(p(x))dx, \tag{10}$$

where the integral being evaluated over the support set of the random variable x. One main difference between discrete and differential entropies is that the differential entropy can become negative. However, we shall see that the differential version of the MI (the difference between two entropies) will always be nonnegative.

In the same way as in the discrete case, conditional entropy and joint entropy can be defined for continuous random variables. The differential MI $I(x;y)$ is given by

$$I(x;y) = h(x) - h(x|y) = \int p(x)\int p(y|x)\log\left(\frac{p(y|x)}{p(y)}\right)dydx. \tag{11}$$

It can be proven that the differential MI has the same properties as in the discrete case.

One practical problem with the definition of the differential MI is the evaluation of the double integral term. Even for Gaussian distributed random variables there exists no closed form solution for (11). In the next section, we will show that (11) can be evaluated under very general assumptions using Monte Carlo methods. Alternatively, the continuous random variables underlying the differential entropy and MI may be quantized. It can be shown that the discrete entropy of an n-bit quantization of a continuous random variable is approximately $h(x) + n$ if $h(x)$ is the continuous entropy [5]. For the MI, it turns out to be even simpler to find a relation between the discrete and the differential versions since

$$\begin{aligned} I(x^{\triangle}; y^{\triangle}) &= H(x^{\triangle}) - H(x^{\triangle}|y^{\triangle}) && (12)\\ &\approx h(x) + n - (h(x|y) + n) && (13)\\ &= I(x;y), && (14)\end{aligned}$$

where $x^{\triangle}$ and $y^{\triangle}$ are the n-bit quantized versions of the continuous random variables x and y, respectively. In other words, for practical considerations one could treat differential MI by using a suitable quantization of the continuous pdf's and evaluating the discrete MI. This relationship might also serve as justification of the discretization of the feature space done in Section 5.

4.3 Monte Carlo Evaluation of Mutual Information

Section 5 describes the quantization of continuous random variables. Here, we turn to the computation of MI by Monte Carlo sampling, which is a standard technique described in many textbooks (see, for example [21], as a statistically oriented one). It has been used recently in other contexts in computer vision [10] and robotics [9]. Equation (11) shows an interesting fact of the MI that can be exploited during evaluation and (11) can be rewritten as

$$I(x;y) = E_{p(x)}\left[E_{p(y|x)}\left[\log\left(\frac{p(y|x)}{p(y)}\right)\right]\right], \tag{15}$$

where we compute the expected value of a random variable twice, first of the random variable $Z_1 = \log\left(\frac{p(y|x)}{p(y)}\right)$ distributed with $p(Z_1) = p(y|x)$ for fixed x and then the expectation of the random variable

$$Z_2 = E_{p(y|x)}\left[\log\left(\frac{p(y|x)}{p(y)}\right)\right]$$

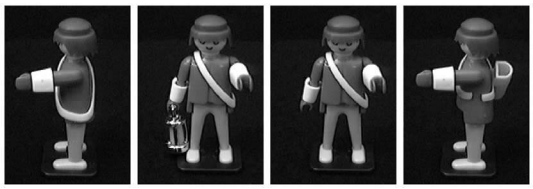


Fig. 4. Four images of three different objects: The first is ambiguous with respect to the objects in image two and three. The second and third view allow for a distinction of objects two and three, but not to distinguish object one from four (the objects with and without quiver on the back).

distributed with $p(Z_2) = p(x)$. The expected value of a random variable $f(Z)$ can be computed by sampling z_i from the distribution $p(Z)$ and computing the mean

$$\widehat{E}_{p(Z)}[f(Z)] = \frac{1}{n}\sum_{z_i} f(z_i) \tag{16}$$

for $1 \leq i \leq n$. The law of large numbers states that $\widehat{E}_{p(z)}[f(z)]$ will converge to $E_{p(z)}[f(z)]$ with probability one [21]. The estimated Monte Carlo standard error of $\widehat{E}_{p(Z)}[f(Z)]$ is

$$\frac{1}{\sqrt{n}}\sqrt{\frac{\sum\left(f(z_i) - \widehat{E}_{p(Z)}[f(Z)]\right)^2}{n-1}}. \tag{17}$$

Under the assumption that one can sample from $p(y|x)$ and $p(x)$ and that both distributions can be evaluated at y and x, respectively, the differential MI in (11) can be approximated using (15) and the Monte Carlo sampling defined in (16).

The assumptions are easily fulfilled by many distributions that occur in computer vision, like Gaussian distributions and even mixtures of Gaussian distributions. Since it is known that any distribution can be approximated by a mixture of Gaussians, the proposition above holds for practically any distribution. Using a mixture of Gaussians for the distributions yields an approach similar to Parzen densities as nonparametric representations of arbitrary densities [23]. We are less interested in a Parzen representation of arbitrary densities and more in the evaluation of the MI for a given continuous pdf, especially of Gaussian distributions used in the next section.

5 Camera Parameter Selection in Object Recognition

If an object recognition system makes its decision based on a single image, ambiguities between objects cannot always be resolved. In the first view of Fig. 4, the unique feature, the lamp in the hand of one of the manikins (see image two and three in Fig. 4), cannot be seen. Depending on the costs for misclassification in such an ambiguous case, either the object should be rejected or a class should be guessed. In any event, taking a second view, where the presence of the lamp can be determined (second and third images), yields a better chance for correct recognition.

Ambiguity is a more serious problem during the design or training of the classifier because such ambiguous views form the difficult examples. Sometimes they cannot be classified correctly even if they are in the training set. Thus, the ultimate goal would be to provide the classifier only with views that are easy to classify. In the context of active object recognition, the MI defines the usefulness of certain views for the following classification step.

In the following experimental section, we look for an optimal camera parameter setting to classify an object. The motivation is that nearly ambiguous objects are easier to classify using their distinguishing subparts. A related idea is to select the best viewpoint (compare Fig. 4). Both ways are tested in the experimental section, i.e., adjusting the (pan, tilt, zoom) settings of the camera (gaze control), as well as, moving the camera on a hemisphere around the object (viewpoint selection).

During a training step for each camera parameter $\boldsymbol{a}_l$, we observe for each object Ω_κ, $\kappa = 1 \ldots K$, a certain feature $\boldsymbol{c}$. The class label Ω_κ can be related to the state $\boldsymbol{x}$ used in Section 3. The feature $\boldsymbol{c}$ is the observation $\boldsymbol{o}$. Obviously, the state $\boldsymbol{x}$ is time invariant in a pure classification problem. Embedded in a statistical context, this means that the pdfs

$$p(\boldsymbol{c}|\Omega_\kappa, \boldsymbol{a}_l) \text{ and } p(\boldsymbol{c}|\boldsymbol{a}_l) \tag{18}$$

can be estimated during training. A common approach is to make some assumptions about the underlying distributions and to estimate the parameters of the distribution. For the estimation, one approach is to choose some or all camera parameters $\boldsymbol{a}_l$ in a supervised learning step. A feature extraction mechanism transforms the image $\boldsymbol{f}_{\boldsymbol{a}_l}$ into a feature $\boldsymbol{c}$. All that matters is that $p(\boldsymbol{c}|\Omega_\kappa, \boldsymbol{a}_l)$ and $p(\boldsymbol{c})$ must be represented (for example, a look-up table in the discrete case, or a parametric function like a Gaussian distribution in the continuous case) and estimated during a training step (for example, estimating the parameters of the Gaussian distribution by a ML estimation) or that these distributions be known by modeling and analysis. One feature we use below, is the mean image gray-level value. This simple scalar feature is easy to extract and learn and it illustrates that even such a weak feature is effective if the camera parameters are chosen using our scheme.

As soon as the densities in (18) have been estimated (in the discrete case, the relative frequency of an observed feature for a given class and action is computed) as already described in Section 3 the MI can be used to decide on the optimal parameters $\boldsymbol{a}_l$ given the a priori probability $p_\kappa = p(\Omega_\kappa)$ of each of the classes Ω_κ. The new camera parameters are used to take a new image. The MI in the notation given above is

$$I(\Omega; \boldsymbol{c}|\boldsymbol{a}_l) = \sum_{\kappa}^{K} \int_{\boldsymbol{c}} p_\kappa p(\boldsymbol{c}|\Omega_\kappa, \boldsymbol{a}_l) \log \frac{p(\boldsymbol{c}|\Omega_\kappa, \boldsymbol{a}_l)}{p(\boldsymbol{c}|\boldsymbol{a}_l)}\, d\boldsymbol{c}, \tag{19}$$

where $\kappa = 1 \ldots K$ is the class label. The value of $I(\Omega; \boldsymbol{c}|\boldsymbol{a}_l)$ is zero if the classes and the features are uncorrelated and reaches its maximum at $-\sum p_\kappa \log p_\kappa$ if each feature can be observed only for exactly one object.

For the experiments in Section 6.1, the range of the feature $\boldsymbol{c}$ is discretized, so that the integration in (19) is reduced to a summation over the discrete values $\boldsymbol{c}_i$

$$I(\Omega; \boldsymbol{c}|\boldsymbol{a}_l) = \sum_{\kappa} \sum_{\boldsymbol{c}_i} p_\kappa p(\boldsymbol{c}_i|\Omega_\kappa, \boldsymbol{a}_l) \log \frac{p(\boldsymbol{c}_i|\Omega_\kappa, \boldsymbol{a}_l)}{p(\boldsymbol{c}_i|\boldsymbol{a}_l)}. \tag{20}$$

We discretized the range of the feature values representing the mean gray value in the image from zero to 255 into eight equally sized intervals. Now, the discrete densities $p(\boldsymbol{c}_i|\Omega_\kappa, \boldsymbol{a}_l)$ and $p(\boldsymbol{c}_i|\boldsymbol{a}_l)$ can be estimated in a training step for each camera parameter setting. The estimation is done by counting the occurrence of pairs of Ω_κ and $\boldsymbol{c}_i$ for a given action $\boldsymbol{a}_l$.

One straightforward way to generalize the tabular representation of the densities is to use a Parzen window density representation and apply the stochastic maximization algorithm EMMA to the maximization of the MI as described in [24], [23]. In Section 4.2, we presented another way to use continuous densities and Monte Carlo evaluation of the MI. It is applied in the experiments in Section 6.2. In the experiments, we also employ a second Bayesian classifier that uses an eigenspace classifier [14], briefly summarized next.

5.1 Statistical Eigenspace Classifier

In contrast to the Bayesian classifier based on the weak feature of the mean gray value and the discrete MI, in Section 6.2, we will also show how we apply the concept of differential MI to view point selection for object recognition. In order to do so, we use a more sophisticated statistical classifier [3], [17] that is derived from an eigenspace approach. The main idea and formalism are summarized here. The eigenspace approach was first introduced in [14]. The key idea is to transform the images interpreted as a row vector of pixel values into a lower dimensional space using principal component analysis (PCA). The mapping $\boldsymbol{\Phi}$ from high dimensional image space $\boldsymbol{f}$ to low dimensional feature space $\boldsymbol{c} = \boldsymbol{\Phi}\boldsymbol{f}$ is defined by computing the eigenvalues of the matrix $\boldsymbol{Q} = \boldsymbol{F}\boldsymbol{F}^T$ with $\boldsymbol{F}$ containing the normalized training images of the different objects from the database. The eigenvectors φ_l that correspond to the k largest eigenvalues of $\boldsymbol{Q}$ then form the matrix

$$\boldsymbol{\Phi} = (\varphi_1, \varphi_2, \ldots, \varphi_k)^T.$$

For the selection of the best (pan, tilt, zoom) setting of the camera defined by the maximum of MI, we need a description of the relationships of object class and image (feature) in a probabilistic framework. This means that we need densities $p(\boldsymbol{c}|\Omega_\kappa)$ for each object class. Although, there exists a very promising approach for probabilistic principal component analysis that results directly in the desired densities [22], for simplicity our implementation follows the approach of [2].

In the following, we assume that, for a given transformation $\boldsymbol{\Phi}$, images $\boldsymbol{f}$ from class Ω_κ are Gaussian distributed in the feature space $\boldsymbol{c}$. In other words, one can define $p(\boldsymbol{c}|\Omega_\kappa)$ by

$$p(\boldsymbol{c}|\Omega_\kappa) = p(\boldsymbol{\Phi}\boldsymbol{f}|\Omega_\kappa) = N(\boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa^{-1}), \tag{21}$$

Maximum-likelihood estimation for the parameters $\boldsymbol{\mu}_\kappa$ and $\boldsymbol{\Sigma}_\kappa^{-1}$ can be done by projecting a large number of test images of object class Ω_κ into the eigenspace using the computed transformation matrix $\boldsymbol{\Phi}$.

In the case of view point selection, the densities $p(\boldsymbol{c}|\Omega_\kappa, \boldsymbol{a})$ can be estimated the same way, i.e., for each (pan, tilt, zoom) setting $\boldsymbol{a}$ of the camera we train a Gaussian distribution

$$p(\boldsymbol{c}|\Omega_\kappa, \boldsymbol{a}) = N_{\boldsymbol{a}}(\boldsymbol{\mu}_\kappa, \boldsymbol{\Sigma}_\kappa^{-1}). \tag{22}$$

A detailed discussion of the original work can be found in [3], [17]. Finally, for n classes and m different (pan, tilt, zoom) settings $\boldsymbol{a}$, we end up with a total number of $m \cdot n$ Gaussian distributions, which are necessary for the computation of the differential MI in (11). In our case, $m = 776$ and $n = 9$.

6 Experimental Results

In Section 6.1, we describe experiments with real camera movements using a discrete density representation and the Bayesian classifier based on the discretized mean gray value as feature. With this feature, it is impossible to classify objects reliably without smart sensor data acquisition. Of course, we are aware of all the obvious problems of this ridiculously weak feature: We chose it exactly because the benefits of our approach can be best shown with a feature that obviously needs smart sensor data selection. To show that strong, modern techniques can also benefit from our approach, in Section 6.2, we present experiments using a statistical eigenspace approach as classifier, continuous densities, and Monte Carlo evaluation of the MI. More experiments with real data and simulated camera parameters appear in [7].

6.1 Parameter Selection Using Discrete Mutual Information

Our data set consists of nine different objects (Fig. 5). Some of the objects have been modified so that they look similar. Two objects (o2 and o5) are so similar that a distinction using the discretized mean gray value as feature is impossible (the central patch is actually a different color). From Fig. 5, it is obvious that, with this impoverished feature, a classification without smart focal length and gaze control is impossible. In fact, the eight-level quantized mean gray value is the same for all the full-resolution overview images shown in Fig. 5.

To perform classification, the following quantities from Section 5 must be specified, where, in contrast to the general case, the state and the observation are scalar values:

- The state x is a discrete class number from 0 to 8.
- The observation o is the mean gray value in the observed image, discretized uniformly to values from 0 to 7.
- The action $\boldsymbol{a} = (p, t, z)^T$, with p, t, and z being the (pan, tilt, zoom) setting of the camera.

Also, these quantities are discrete values. For the zoom position, six discrete values have been chosen, resulting in a range between overview and close-up view, indicated in Fig. 6. The range of pan and tilt is dependent on the selected focal length to avoid imaging the background.

During training, the different densities in (19) must be estimated. The most important part is the estimation of the conditional density $p(o|x, \boldsymbol{a})$. Thus, for all objects in a

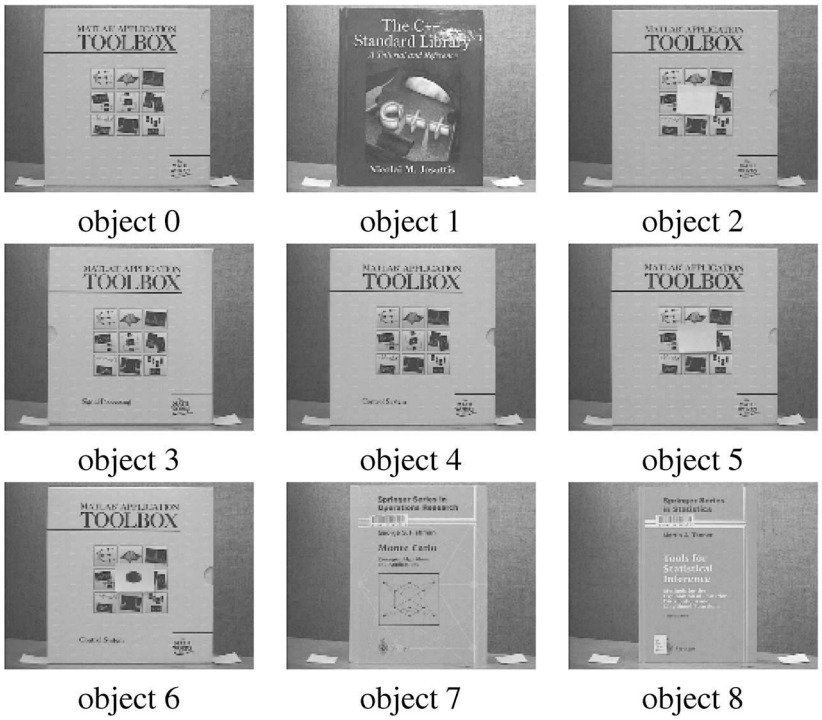


Fig. 5. Data set for classification using zoom planning.

supervised step, different parameters for the camera are set and the feature is extracted from the resulting image. While repeating this a sufficient number of times (in the experiments each (pan, tilt, zoom) setting was set for each object between 100 and 10,000 times to check the influence of the number of training examples during training on the selected camera parameter during test), the density $p(o|x, \boldsymbol{a})$ can be estimated by computing the relative frequency of the observed feature o.

The experiments were performed as follows (compare Fig. 3):

1. **Initialization.** The distribution over the nine classes are initialized uniformly, to take into account that a priori (and from the overview images) no information favoring any class is available.
2. **Parameter selection.** Based on the a priori probability, the best (pan, tilt, zoom) setting is computed using the MMI criterion (7).
3. **Imaging and feature extraction.** The (pan, tilt, zoom) settings are commanded to the camera. An image is taken and the feature (quantized mean gray value) is extracted.

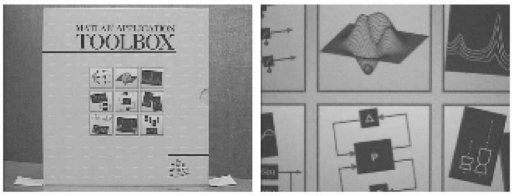


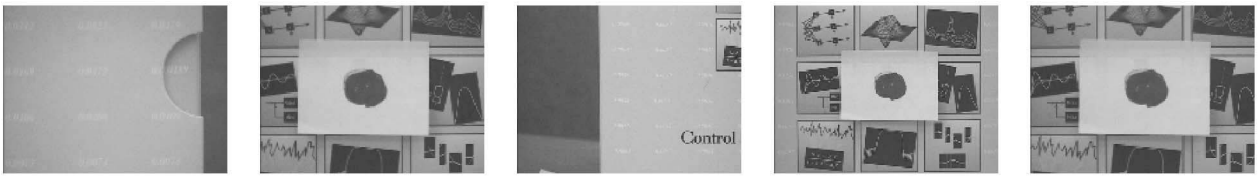
Fig. 6. Range in focal length: Left, shortest focal length. Right, longest focal length.

4. **Bayes decision.** Bayes formula is used to compute the a posteriori probability for the nine classes.
5. **Loop or end.** If the a posteriori probability for one class is greater than 0.9 (an arbitrary constant) or 10 views (another arbitrary constant) have been already taken, then end. Else, set the a priori probability for the next time step to the current a posteriori probability. Go to 2.

The reader must remember that the information used by the automatic process is simply one of eight scalar integer numbers—the quantized mean gray value of the image.

Fig. 7 depicts a typical experiment. Several more can be found in [7]. Besides the change in belief state for the nine classes, one can also see the change in entropy of the distribution over the classes. Except for the transition from view 2 to view 3, the entropy is reduced step by step, which finally results in a unique and correct decision for object number 6. The increase in entropy can be explained by an error in the noise model, i.e., the true noise has been underestimated in this case. Nevertheless, the sequential decision process results in the correct classification.

Fig. 7 is also a good example to show that the system has learned to look at the important parts of the objects. After the first selected view, it can exclude object 1, 7, and 8 from the hypotheses set. Then, only the Matlab boxes are possible hypotheses and, therefore, the center of the boxes contains the most information at the next time step. This part is focused on during the next time interval, as can be seen in Fig. 7, view 2 (top row, second image). The reason for the repeated, identical look to the center (view 2 and view 5) can be explained by a mismatch between the learned and the true underlying model for the objects. One can observe that the entropy after selected view 2 increases. Also, the maximum a posteriori probability would return object



view 1 view 2 view 3 view 4 view 5

view	o0	o1	o2	o3	o4	o5	o6	o7	o8	entr.
initial	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.111	0.95
view 1	0.251	0.000	0.251	0.000	0.073	0.251	0.095	0.078	0.000	0.72
view 2	0.523	0.000	0.034	0.000	0.113	0.014	0.256	0.050	0.000	0.50
view 3	0.125	0.000	0.069	0.000	0.237	0.027	0.542	0.000	0.000	0.53
view 4	0.003	0.000	0.092	0.000	0.000	0.043	0.861	0.000	0.000	0.22
view 5	0.003	0.000	0.005	0.000	0.000	0.000	0.990	0.000	0.000	0.03

Fig. 7. Top: object recognition using gaze control for object number 6 (the initial overview has been left out). Bottom: the corresponding a posteriori probability over the objects number 0 to 8 and the resulting entropy for the views 1 to 5 from the top figure .

number 0 as the classification result. During the next verification steps, the system comes back to the right decision, i.e., maximum a posteriori probability for object number 6. And to finally certify this result, another gaze to the center (view 5 in Fig. 7) is necessary.

With higher noise in the camera parameter control, the result is that the entropy never increases [7], but the decrease in uncertainty is dramatically reduced and, also, the final decision is not as unambiguous as for the experiment shown in Fig. 7. Regardless, the maximum a posteriori decision after the last view returns the right class, i.e., object number 6.

Table 1 gives the recognition results for the nine objects. In the first row, the noise in the camera movement and focal length adjustment has been assumed to be low; in the second row it has been assumed to be high. Actually, the true noise in the control of the camera parameter is unknown and has not been estimated for this work. The last row shows the results for random gaze and focal length control for selected objects. Objects 2 and 5 could not be distinguished based on the mean gray value (compare also Fig. 5), Thus, both objects are considered as one class that is distinguished from the other seven classes. As expected, assuming more noise in the camera control the system will less often choose a close-up view, which results in a reduced total recognition rate, although the easier objects (1, 7, and 8) can be recognized as well as or even more reliably compared to the experiments with an optimistic noise assumption. Comparing the results with a random gaze control (third row in Table 1) for objects 1, 2, and 6, one can conclude the following: For the easy recognizable object 1, a random strategy results in the same recognition rate, although the mean number of views is increased from 1 to 2.5 views (compare Table 2). For object 2, which is more complicated to recognize reliably, one gets an error of 30 percent compared to zero error using the proposed sequential decision process. Finally, object 6 is an example where the random strategy practically fails completely with an error rate of 80 percent.

6.2 Parameter Selection Using Differential Mutual Information

In this section, we present experiments with the statistical eigenspace classifier and differential MI. First, we use same data set as before (Fig. 5). Then, we present experiments with a different data set and viewpoint selection actions.

In the training step for each (pan, tilt, zoom) setting $\boldsymbol{a}$, we took views from each object class Ω_κ to compute the transformation matrix $\boldsymbol{\Phi_a}$. Afterwards, we synthetically created a total number of 100 new disturbed views for each object class and projected the images into the eigenspace. The disturbance during this training step is a random shift in x and y position of the captured window as well as pixelwise additive Gaussian noise with a variance of $\sigma^2 = 15$. The noise components model inaccuracies in camera positioning and noisy image formation. The resulting feature vectors $\boldsymbol{c}_i$ are used for a maximum-likelihood estimation of the parameters of the Gaussian densities.

TABLE 1
Recognition Results (in Percent)

exper	o0	o1	o2	o3	o4	o5	o6	o7	o8	total
low noise	80	100	100	0	80		90	70	100	77.5
high noise	0	100	60	100	40		50	100	100	68.7
random		100	70				20			

First row, a low noise assumption; second row, a high noise assumption; third row (for selected objects), a random strategy.

TABLE 2
Average Number of Views until Decision

exper	o0	o1	o2	o3	o4	o5	o6	o7	o8	total
less noise	4.7	1	10	10	4.3	0	5	2	2	4.9
more noise	10	2	10	10	10	0	10	10	3	8.1
random		2.5	10				10			

First row, low noise assumption; second row, high noise assumption; third row, (for selected objects) a random strategy.

During the test, we compared the sequential decision process again with a random strategy. The procedure is the same as already described earlier. The main differences with the Bayesian classifier using the mean gray value is that now continuous probability densities are used together with differential MI for selection of the best next (pan, tilt, zoom) setting $\boldsymbol{a}$ and that a statistical Eigenspace approach for classification is applied.

Table 3 gives the results for the view point planning strategy based on the maximum of MI. The decision for the next view is made by Monte Carlo evaluation (with 1,000 samples) of the MI as described in Section 4.3. Almost all objects could be recognized perfectly although the number of views necessary for the decision varies between the different classes. For example, the objects o0, o2, o3, o4, o5, and o6 are the difficult cases since these objects look very similar. This similarity is expressed in the results by an increased mean number of views necessary for recognition. However, the recognition rate still is 100 percent or close to it.

It is natural that object o1 is recognizable in any case in the first view. It is interesting to look at the maximum a posteriori probability that results after the right decision has been made. Again, in almost every case, the maximum a posteriori probability is greater than 0.95, which corresponds to a very certain decision for the right class or—in other words—in a small entropy for the a posteriori probability.

In comparison to the random strategy shown in Table 3, the maximum a posteriori probability is much less than 0.9 in the case of a correct decision. As a consequence, the decision is more uncertain. Also, the recognition rate is dramatically reduced (with the exception of objects o1, o7, and o8). In most cases, the full number of 10 trials is made after which the decision is finally forced.

Although, the recognition rate for the "easy" objects—o1, o7, and o8—is comparable to the results using view point planning, the mean number of views that are necessary to return an a posteriori probability of more than 0.9 is almost twice as large for object o7 and o8. Object o1 turned out to be recognizable quickly and robustly in either case, although a marginal difference exists in the overall results for recognition rate and mean number of views.

The total recognition rate is improved from 81.4 percent for random strategy to 99.8 percent for with gaze planning. Our gaze selection strategy based on MMI works in practice for a standard, state of the art classification method and outperforms a random strategy.

Finally, we did experiments with active viewpoint selection [3], [17]. Five toy manikins form the data set (three of them are shown in Fig. 4). There are only certain views from which the objects can be distinguished. The experimental setup consists of a turntable and a robot arm with a camera mounted that can move around the turntable. The experimental setup is described in more detail in [6]. The actions $\boldsymbol{a} = (\phi, \theta)^T$ define the position of the camera on the hemisphere. Again, the statistical eigenspace approach is used for the classifier.

Table 4 summarizes the results. As before, the planning based on MMI outperforms a random strategy. However, the gain in recognition rate is less than in the case of gaze control since the object are less ambiguous. In fact, in most cases, the object can be recognized with three views at the latest. In Fig. 8 (left), the MI (from (19)) is shown at the beginning of the sequential decision process, i.e., the prior is

TABLE 3
Results for Gaze Planning and Random Gaze Control (1,000 Trials per Object)

	planned gaze control			random gaze control		
object	rec. rate	mean no. views	mean max. prob.	rec. rate	mean no. views	mean max. prob.
o0	99.5	2.4	0.96	83.4	9.9	0.61
o1	100.0	1.0	1.00	99.6	1.2	1.00
o2	100.0	4.0	0.95	62.4	9.8	0.65
o3	100.0	2.3	0.96	76.0	9.8	0.64
o4	100.0	4.0	0.95	66.6	10.0	0.56
o5	99.2	3.5	0.97	68.2	9.9	0.57
o6	99.6	2.8	0.96	76.7	9.7	0.63
o7	100.0	1.7	0.98	100.0	2.5	0.97
o8	100.0	1.1	1.00	100.0	2.4	0.97
average	99.8	2.5	0.97	81.4	7.2	0.73

Recognition rate, mean number of views, and maximum a posteriori probability for the right class after the decision has been made.

TABLE 4
Results for Viewpoint Planning and Random Viewpoint Control (100 Trials per Object)

object	planned viewpoint control			random viewpoint control		
	rec. rate	mean no. views	mean max. prob.	rec. rate	mean no. views	mean max. 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
trupet	99	2.29	0.97	70	8.89	0.65
average	94.2	1.56	0.97	83.0	6.03	0.84

Recognition rate, mean number of views, and maximum a posteriori probability for the right class after the decision has been made.

assumed to be uniform. The x- and y-axis are the motorsteps for moving the turntable and the robot arm, to define views on the hemisphere. The motorstep values correspond to a rotation between 0 and 360 degrees for the turntable and -90 to 90 degree for the robot arm. The MI has been computed by Monte Carlo simulation as described in Section 4.2. The maximum in this 2D function in the case of viewpoint selection defines the best action (viewpoint) to be chosen. In Fig. 8 (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.

The computation time for the computation of the best action depends in the case of a discrete action space on the number of actions $\boldsymbol{a}$, the number of discrete features and the number of classes. In the case of the differential MI, the computation time depends on the number of actions, the number of classes, and the number of samples taken to approximate the differential MI. In practice, for optimal gaze selection less than one second is needed on a Pentium II/300 for the computation of the best action using 1,000 samples, nine classes, and a total of 776 different actions.

7 Conclusion

State estimation is a formalism that can be used to frame the most important problems in computer vision. Clearly, the observations (images, features, high-level structures) have a strong influence on the accuracy of state estimation. Thus, either implicitly or explicitly most systems cycle through a state estimation and action selection stage. In the paradigm of active vision, it remains an unsolved problem in general which sensor data should be selected at a certain stage of state estimation.

Rather than optimizing an estimator-specific metric (building a better edge-finder or classification algorithm), we desire a general way to reduce the uncertainty in any state estimation process using estimator independent techniques. The main assumption is that every state estimator will return better results if the uncertainty in the state estimation process is reduced in advance. This separation of our process from a particular state estimator makes our approach most general and independent from the state estimator at hand. While we do not (currently) improve the state estimator, we do provide the state estimator with the best sensor data at each decision stage.

To measure the uncertainty in the state estimation process, we have introduced a formalism based on Shannon's information theory. The important quantity in our work is the conditional MI, conditioned on the selected camera parameters. The MI between the distributions over the state and the observations measures how much information the observation will contain about the state, or, in other words, how much uncertainty about the state is reduced by collecting observations. As a consequence, maximizing the conditional MI with respect to the controllable information-acquiring actions returns the best action in terms of reduction in uncertainty.

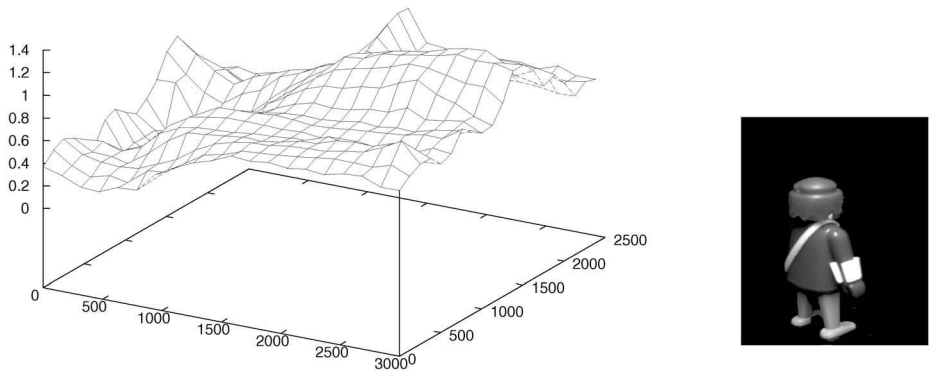


Fig. 8. Left: MI in the viewpoint selection example assuming a uniform prior (computed by Monte Carlo evaluation). The x and y are the motorsteps for the turntable and robot arm, respectively. Along the z, the MI is ploted. Right: best view $\boldsymbol{a}$ decided by the maximum in the MI, depicted left ($\boldsymbol{a} = (2, 550, 1, 500)$).

To show the quality and problems of our approach, we used an object recognition scenario, i.e., a state estimation problem of a static system using active gaze control. In the experiments, our approach was able to achieve a recognition rate of more than 77 percent despite a very weak feature and a very difficult data set. Without active sensor data selection, the objects could not be classified at all. Also, our approach outperforms a random strategy for action selection in both the number of views necessary for classification as well as in the recognition rate. We show similar results with a more sophisticated statistical eigenspace classifier. The camera parameter selection strategy based on the differential mutual information (recognition rate, 99.8 percent) again outperforms the random strategy (recognition rate: 81.4 percent). The higher overall recognition rate is due to the better features extracted in the eigenspace approach. Similar results could be reported for a view point selection scenario.

The benefits of our approach lie in the systematic reduction of uncertainty about the true state by selecting an optimal sequence of actions and the independence from the state estimator. Another important result is that the convergence of the sequential decision process can be proven. The approach can be combined with any state estimator that fulfills the following assumptions: First, the unobservable, true state is estimated using observations that are correlated with the true state. Second, the state estimator returns an a posteriori probability distribution over the state space. Last, the conditional pdf's (conditioned on the action) for the observations and the likelihood function must be known or estimated in a training step. These three assumptions are met by many if not by most of the state estimators used in computer vision.

Our approach is completely embedded in a statistical framework and the estimation of the parameters of the densities is not a trivial problem, especially in higher dimensional spaces (state, feature, and action). So far, we do not optimize or adapt the parameters of the state estimator. As a consequence, the sequential decision process will not improve state estimation if the state estimator systematically returns wrong or strongly biased state estimates. A quite natural idea would be to look for an integration of this sequential decision process into a framework that allows the optimization of the state estimator itself by changing its parameters. One promising starting point for such an integration of our work with approaches from state estimation is the work on active learning [4].

In our future work, we will apply a more general approach for representing pdf's of random vectors, the so-called Parzen window density estimation. In [24], [23], an approach, EMMA, has been developed for maximizing the MI of two random variables represented by a Parzen window density for alignment of images of different modalities. Such an algorithm for maximization of the MI could be directly relevant when we extend the discrete actions space to a continuous one. Finally, we are working on extending our framework to state estimation in dynamic systems.

ACKNOWLEDGMENTS

The authors thank J. Triesch and R. Jacobs for discerning reviews of the manuscript and B. Madden for his detailed comments and discussions. They also would like to thank the three anonymous referees for suggesting several good ways to improve the original manuscript. This research was supported by grant DE 735/1 of the German Science Foundation (DFG) and partially by the US National Science Foundation grant EIA-9972881 and CAT/NYSSTF grant EEC-9813002.

REFERENCES

[1] T. Arbel and F.P. Ferrie, "Viewpoint Selection by Navigation through Entropy Maps," *Proc. Seventh Int'l Conf. Computer Vision,* 1999.

[2] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "Active Object Recognition in Parametric Eigenspace," *Proc. British Machine Vision Conf.,* vol. 2, pp. 629-638, 1998.

[3] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz, "Appearance Based Active Object Recognition," *Image and Vision Computing,* vol. 18, pp. 715-727, 2000.

[4] D.A. Cohn, A. Ghahramani, and M.I. Jordan, "Active Learning with Statistical Models," *J. Artificial Intelligence Research,* vol. 4, pp. 129-145, 1996.

[5] T.M. Cover and J.A. Thomas, *Elements of Information Theory.* New York: Wiley and Sons, 1991.

[6] F. Deinzer, J. Denzler, and H. Niemann, "Viewpoint Selection—A Classifier Independent Learning Approach," *Proc. IEEE Southwest Symp. Image Analysis and Interpretation,* pp. 209-213, 2000.

[7] 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 Dept., Univ. of Rochester, 2000.

[8] J. Fisher and J.C. Principe, "A Nonparametric Method for Information Theoretic Feature Extraction," *Proc. Defense Advance Research Projects Agency (DARPA) Image Understanding Workshop,* 1997.

[9] D. Fox, W. Burgard, and S. Thrun, "Active Markov Localization for Mobile Robots," Technical Report, Carnegie Mellon Univ., 1998.

[10] M. Isard and A. Blake, "Contour Tracking by Stochastic Propagation of Conditional Density," *Computer Vision—ECCV '96,* A. Blake, ed., pp. 343-356, 1996.

[11] L.P. Kaelbling, M.L. Littman, and A.R. Cassadra, "Planning and Acting in Partially Observable Stochastic Domains," *Artificial Intelligence,* vol. 101, nos. 1-2, pp. 99-134, 1998.

[12] R.E. Kalman, "A New Approach to Linear Filtering and Prediction Problems," *J. Basic Eng.,* pp. 35-44, 1960.

[13] J.M. Manyika and H.F. Durran-Whyte, "On Sensor Management in Decentralized Data Fusion," *Proc. Conf. Decision and Control,* pp. 3506-3507, 1992.

[14] H. Murase and S. Nayar, "Visual Learning and Recognition of 3-D Objects from Appearance," *Int'l J. Computer Vision,* vol. 14, pp. 5-24, 1995.

[15] H. Niemann, *Pattern Analysis and Understanding.* vol. 4, Berlin, Heidelberg: Springer, 1990.

[16] C.A. Noonan and K.J. Orford, " Entropy Measures of Multi-Sensor Fusion Performance," *Proc. IEE Colloqium Target Tracking and Data Fusion,* pp. 15/1-15/5, 1996.

[17] L. Paletta and A. Pinz, "Active Object Recognition by View Integration and Reinforcement Learning," *Robotics and Autonomous Systems,* vol. 31, pp. 71-86, 2000.

[18] L. Paletta, M. Prantl, and A. Pinz, "Learning Temporal Context in Active Object Recognition Using Bayesian Analysis," *Proc. Int'l Conf. Pattern Recognition,* vol. 3, pp. 695-699, 2000.

[19] B. Schiele and J.L. Crowley, "Transinformation for Active Object Recognition," *Proc. Sixth Int'l Conf. Computer Vision,* 1998.

[20] R.S. Sutton and A.G. Barto, *Reinforcement Learning.* Cambridge, London: Bradford, 1998.

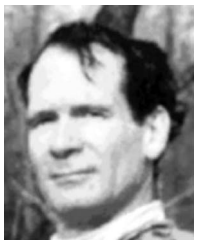
[21] M.A. Tanner, *Tools for Statistical Inference.* London: Springer Verlag, 1993.

[22] M.E. Tipping and C.M. Bishop, "Mixtures of Probabilistic Principal Component Analysers," *Neural Computation,* vol. 11, no. 2, pp. 443-482, 1999.

[23] P. Viola and W.M. Wells III, "Alignment by Maximization of Mutual Information," *Int'l J. Computer Vision,* vol. 24, no. 2, pp. 137-154, 1997.

[24] P.A. Viola, "Alignment by Maximization of Mutual Information," AI Technical Report No. 1548, MIT Artificial Intelligence Lab., 1995.
[25] D. Wilkes, "Active Object Recognition," Technical Report RBCV-TR-94-45, Dept. Computer Science, Univ. of Toronto, 1994.
[26] Y. Ye, "Sensor Planning for Object Search," PhD thesis, Dept. Computer Science, Univ. of Toronto, 1997.

Joachim Denzler received the Diplom-Informatiker degree from University Erlangen-Nuremberg, Germany, in 1992, in computer science, especially pattern recognition, speech recognition, and theoretical electrical engineering. He received the PhD degree in computer science in 1997. Since January 1993, Dr. Denzler has been a member of the research staff of the chair for pattern recognition of the University of Erlangen. Currently, he holds the position of a research associate. His research activities concentrate on probabilistic modeling of sensor data and action sequences in the field of computer vision. Currently, his main emphasis is concentrated on the optimal selection of camera parameters during classification and tracking of moving objects on the statistically modeled view point selection for optimal object recognition and pose estimation, as well as, on active knowledge based self localization and scene exploration for service robots. Dr. Denzler is a member of the IEEE, the IEEE Computer Society, and Gesellschaft for Informatik.



Christopher M. Brown received the BA degree from Oberlin in 1967 and the PhD degree from the University of Chicago in 1972. He is a professor of computer science at the University of Rochester, where he has been since finishing a postdoctoral fellowship at the School of Artificial Intelligence at the University of Edinburgh in 1974. He is coauthor of *COMPUTER VISION* with his Rochester colleague Dana Ballard. His current research interests are computer vision and robotics, integrated parallel systems performing animate vision (the interaction of visual capabilities and motor behavior), and in the integration of planning, learning, sensing, and control.