

An Information Theoretic Approach to Optimal Sensor Data Selection for State Estimation

J. Denzler C. Brown

Abstract

In this paper we introduce a formalism for optimal sensor parameter selection for iterative state estimation in static systems. In contrast to common approaches, where a certain metric — for example, the mean squared error between true and estimated state — is optimized during state estimation, in this work the optimality is defined in terms of reduction in uncertainty in the state estimation process. The main assumption is that state estimation becomes more reliable if the uncertainty and ambiguity in the state estimation process can be reduced.

We use Shannon's information theory to select the camera parameters that maximize mutual information, thus optimizing the information that the captured image 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 pan/tilt/zoom camera. During the sequential decision process the camera looks to parts of the object that allow the most reliable discrimination between similar objects. We

performed experiments with discrete density representation as well as with continuous densities and Monte Carlo evaluation of the mutual information. The sequential decision process outperforms a random gaze control, both in the sense of recognition rate and number of views necessary to return a decision.

1 Introduction

The state, or state vector, of a system describes the relevant system parameters to be determined from observations by sensors. This paper tackles the problem of optimal sensor data selection for state estimation in computer vision from an information theoretic point of view. 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 which 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, like objects identical in some views (compare Figure 4).

In contrast to classical and modern approaches for state estimation [13, 4] in our approach we do 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 2.

Uncertainty in state estimation will in general increase the variance of the pdf over the state space (sometimes called belief state [16]), while ambiguity increases the number of its modes. Our claim is that uncertainty and ambiguities can be minimized by using the right choice of sensor data. The general principle and goal of our work is depicted in Figure 1. A sequence of actions \mathbf{a}_t is chosen in order to transform a prior distribution $p(\mathbf{x}_t)$ over the state space $\mathbf{x}_t \in \mathbb{R}^n$ ($p(\mathbf{x}_t)$ 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 actions represents any influence on the image acquisition process. In the case of a static system the true state remains constant over time. In the case of a dynamic system the problem of state estimation becomes more difficult, since 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 will focus in the following on static state estimation.

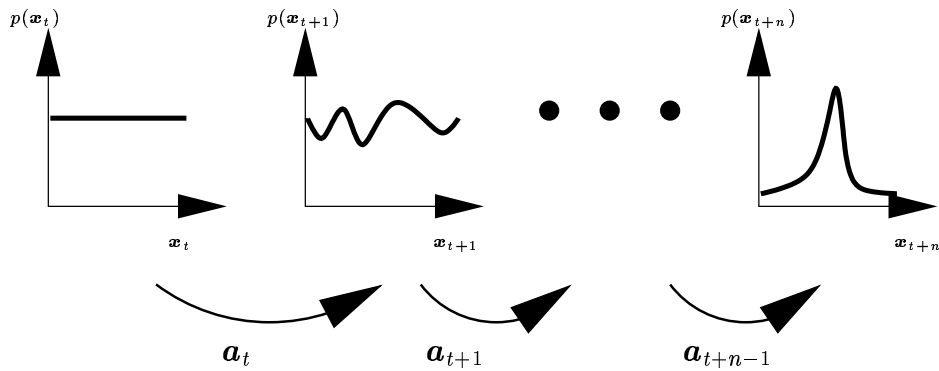


Figure 1: General principle: reduce uncertainty and ambiguity (variance and multiple modes) in the pdf of the state \mathbf{x}_t by choosing appropriate actions \mathbf{a}_t , or to be more precise, by selecting the right sensor data.

In the following we look at one special class of camera actions \mathbf{a}_t , the adjustment of the focal length and the pan/tilt position of a camera — although the framework can be used for any other actions, e.g. iris control or tuning of the focus of the camera. In order to demonstrate the benefits of the approach one important computer vision problem is discussed: object recognition using active camera parameter selection. In the object recognition example 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. Therefore a criterion must be provided that balances this trade-off between long focal length for detailed inspection and short focal length for global overview based on the current information on the state of a static system.

The paper is structured as follows. In Section 2 a formal statement of the problem is given. The next section considers a sequential decision process in the case of a time invariant system, namely in the case of object recognition. The discrete density representation is extended in Section 4 to continuous densities and Monte Carlo evaluation of the mutual information. Also, a more sophisticated classifier using statistical eigenspace is introduced. Related work, applying information theoretic concepts in computer vision, is discussed in Section 5. The experimental evaluation is summarized in Section 6. The paper concludes with a discussion of the results achieved and the problems observed, as well as with a perspective on future work.

2 Formal Problem Statement

2.1 The Probabilistic Framework

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 \mathbf{o}_t or some

other acquired sensor information at time step t the unobservable true state \mathbf{x}_t of the system, either a static or time varying one, is approximated by a state estimate $\hat{\mathbf{x}}_t$. This estimated state is the basis for selecting a certain action \mathbf{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.

In the following we have chosen a probabilistic framework, motivated by the fact that sensor data is not noiseless or ideal, nor can the effect of a certain action be completely determined in advance. In a probabilistic framework this uncertainty can be modeled by adding a stochastic noise component to the parameters that must be estimated. The noise estimation can be done during training, or by making some assumptions, which are verified later on during the working stage of the system.

The probabilistic framework also allows us to use pdf's to describe the current state estimate, instead of deciding on exactly one true state estimate. The distribution is sometimes called the belief state, since it expresses the belief of being in a certain state. In recent years the belief state representation has been extensively used and studied in the context of particle filters [11].

Two examples can be given to clarify our scenario. In object tracking the (time-varying) state of the system could be the position, velocity and acceleration of the object in 3-D and an action would be the selection of a pan and tilt movements 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 new viewpoints that help to increase recognition when some of the views of different objects are ambiguous and cannot be used to decide for a single class with high certainty [6, 3, 17].

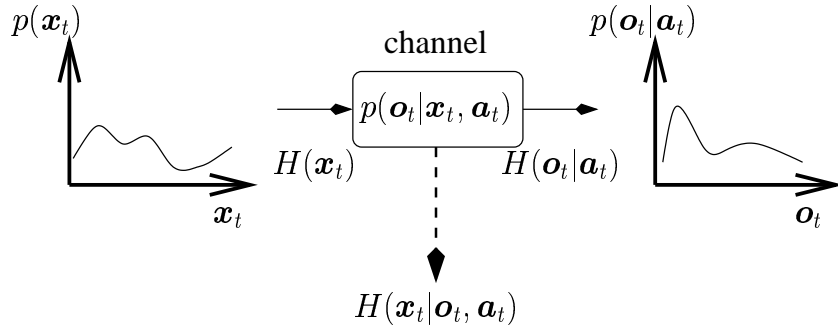


Figure 2: Input and output relation in the channel model and some of the important entropies H describing the information content. The state estimator that estimates the belief state \mathbf{x}_t based on the observation \mathbf{o}_t is missing in this figure.

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

$$p(\mathbf{x}_t | \mathbf{o}_{t-1}, \dots, \mathbf{o}_0) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_{t-1} | \mathbf{o}_{t-1}, \dots, \mathbf{o}_0) p(\mathbf{x}_t | \mathbf{x}_{t-1}) d\mathbf{x}_{t-1} \quad (1)$$

using a model $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ of the dynamics of the system. The density in (1) is often called a temporal prior. As already mentioned the a priori probability is abbreviated with $p(\mathbf{x}_t)$ in Figure 2.

With that pdf an entropy

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

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

$$p(\mathbf{o}_t|\mathbf{a}_t) = \int_{\mathbf{x}_t} p(\mathbf{o}_t|\mathbf{x}_t, \mathbf{a}_t)p(\mathbf{x}_t)d\mathbf{x}_t \quad . \quad (2)$$

Again, an entropy $H(\mathbf{o}_t|\mathbf{a}_t)$ can be associated with the distribution $p(\mathbf{o}_t|\mathbf{a}_t)$. The important quantity in this formalism is the chosen action \mathbf{a}_t . Since the likelihood function $p(\mathbf{o}_t|\mathbf{x}_t, \mathbf{a}_t)$ is conditioned on this action, the action itself influences the properties of the channel. For example, an optimal action \mathbf{a}_t^* would result in a noiseless channel, i.e.

$$p(\mathbf{o}_t|\mathbf{x}_t, \mathbf{a}_t^*) = \begin{cases} 1 & \text{if } \mathbf{o}_t = \mathbf{x}_t \\ 0 & \text{otherwise} \end{cases} \quad . \quad (3)$$

Still, the goal is to estimate the true state \mathbf{x}_t , given the observation \mathbf{o}_t . In information theory an important quantity is used to define how much uncertainty is reduced in \mathbf{x}_t if the observation \mathbf{o}_t

is made. This quantity is called *mutual information* or *transinformation*. In our case, since the information flow through the channel depends on the parameter \mathbf{a}_t we need to define conditional mutual information as

$$I(\mathbf{x}_t; \mathbf{o}_t | \mathbf{a}_t) = H(\mathbf{x}_t) - H(\mathbf{x}_t | \mathbf{o}_t, \mathbf{a}_t) \quad . \quad (4)$$

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

$$I(\mathbf{x}_t; \mathbf{o}_t | \mathbf{a}_t) = \int_{\mathbf{x}_t} \int_{\mathbf{o}_t} p(\mathbf{x}_t) p(\mathbf{o}_t | \mathbf{x}_t, \mathbf{a}_t) \log \left(\frac{p(\mathbf{o}_t | \mathbf{x}_t, \mathbf{a}_t)}{p(\mathbf{o}_t | \mathbf{a}_t)} \right) d\mathbf{o}_t d\mathbf{x}_t \quad . \quad (5)$$

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 mutual information. Since the mutual information is a function of the parameter \mathbf{a}_t the optimal action \mathbf{a}_t^* that can be chosen, given an a priori distribution $p(\mathbf{x}_t)$ and a model for the channel noise $p(\mathbf{o}_t | \mathbf{x}_t, \mathbf{a}_t)$, is defined by

$$\mathbf{a}_t^* = \operatorname{argmax}_{\mathbf{a}_t} I(\mathbf{x}_t; \mathbf{o}_t | \mathbf{a}_t) \quad . \quad (6)$$

An interpretation of this criterion for an optimal action selection is the following. In reality we have more or less reliable state estimators (e.g. object recognizers). We are not interested in optimizing some specific metric corresponding to a certain algorithm, like for example the distance between the true and estimated state of a moving object. The main goal is to reduce uncertainty and ambiguity in the whole process (Figure 1). The assumption is that the measure of mutual information tells us which sensor data must be chosen to make the estimation of the state of a system most reliable.

One problem in statistical approaches is the selection of the various densities involved in equation (5). A common approach is to select a parametric form of the densities and to estimate the parameters during a so-called training step. Due to the curse of dimensionality this can become a very difficult problem. An advantage, though, of the statistical approach is that if the statistical properties of some or all relevant quantities are known, then these distributions may be incorporated directly.

2.2 Sequential Decision Making

We now look at an iterative application of the mutual information framework in state estimation. This iterative procedure distinguishes our approach from other work using mutual information in state estimation (see Section 5).

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

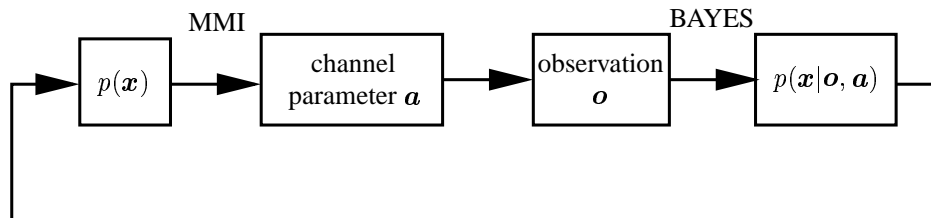


Figure 3: Sequential decision process of maximum mutual information (MMI) for camera parameter selection and Bayesian update of $p(\mathbf{x}|\mathbf{o}, \mathbf{a})$ based on the observed feature \mathbf{o} .

At the beginning of the sequential decision process (let's say at time $t = 0$) the a priori probability over the state space $p(\mathbf{x}_0)$ is initialized uniformly. If reliable, non-uniform priors are known, of course we could use them at this point. The first camera parameter \mathbf{a}_0 is selected based on the maximization of the mutual information (6). Section 3 and 4 show how discrete and differen-

tial mutual information measures are applied. The resulting image (using camera parameter \mathbf{a}_0) or some information extracted from this image serves as observation \mathbf{o}_0 . Bayes rule returns the following a posteriori probability:

$$p(\mathbf{x}_0|\mathbf{o}_0, \mathbf{a}_0) = \frac{p(\mathbf{o}_0|\mathbf{x}, \mathbf{a}_0)p(\mathbf{x}_0|\mathbf{a}_0)}{p(\mathbf{o}_0|\mathbf{a}_0)} = \frac{p(\mathbf{o}_0|\mathbf{x}, \mathbf{a}_0)p(\mathbf{x}_0)}{p(\mathbf{o}_0|\mathbf{a}_0)} \quad . \quad (7)$$

The last step in equation (7) is 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 generation step, i.e. $p(\mathbf{x}_1) = p(\mathbf{x}_1|\mathbf{o}_0, \mathbf{a}_0)$. As a consequence the mutual information in equation (5) will be changed after the first update of the state estimate. In general after the n th view planning step one gets as prior probability of time step $n + 1$

$$p(\mathbf{x}_{n+1}) = \frac{p(\mathbf{o}_n|\mathbf{x}, \mathbf{a}_n)p(\mathbf{x}_n|\mathbf{o}_{n-1}, \dots, \mathbf{o}_0)}{p(\mathbf{o}_n|\mathbf{a}_n)} \quad (8)$$

and

$$\mathbf{a}_n = \operatorname{argmax}_{\mathbf{a}} I(\mathbf{x}_n; \mathbf{o}|\mathbf{a}) \quad . \quad (9)$$

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.

A classical way to describe a sequential decision process is as a Markov decision process. Dynamic programming is the technique that is the basis of most algorithms for configuring Markov decision processes from examples (see for example the textbook on reinforcement learning by

Sutton [18]). Recently the partially observable case [12] 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 (eq. (6)) is not a trivial task. Ideally, this estimation could be unnecessary if such knowledge is provided by the state estimator.

2.3 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 proved [7]. One consequence of the proof is that under certain assumptions that are difficult to verify 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 mutual information 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 mutual information means minimizing the conditional entropy (compare eq. 4). This follows directly from the definition of mutual information. As a consequence one cannot assure that for one single step in the sequential decision process the uncertainty is reduced. The change in uncertainty depends on the current observation. On average, though, i.e. in the long run, by definition of the mutual information the uncertainty will be reduced.



Figure 4: Three images of two different objects: the first view is ambiguous, the second and third allow for a correct classification.

3 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 Figure 4, the unique number on each cup, which is the only difference between the two cups, 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 number can be seen, will yield a higher chance for a 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. The question and main problem is how can we identify such views automatically. Our approach is to use a criterion that defines the usefulness of certain views, and to take those that give the most information for the following classification step.

In the next sections we look for an optimal camera setting to classify an object. The motivation is that difficult objects in the sense of ambiguities are more easy to classify if one does not look at the object as a whole, but instead inspects certain parts of the object. The inspection is done by adjusting the focal length or gaze of the camera. Again there is the trade-off between a global

inspection, which might allow successful classification of the unambiguous objects, and a detailed inspection, which might not be helpful until some objects are ruled out.

The idea is to define the optimal choice of the camera parameters as a feature selection problem in classification. Let us assume that the object to be classified lies in front of a pan/tilt camera, with the optical axis hitting the center of the object. The camera parameter vector \mathbf{a}_l contains the focal length and the position on the object the camera is looking at, coded as the pan/tilt position of the camera. These positions are measured with respect to the zero position (the pan/tilt position where the optical axis hits the center of the object.)

During a training step for each camera parameter \mathbf{a}_l we observe for each object Ω_κ , $\kappa = 1 \dots K$, a certain feature \mathbf{c} . The class label Ω_κ can be related to the state \mathbf{x} , used in Section 2. The feature \mathbf{c} is the observation \mathbf{o} . Obviously the state \mathbf{x} is time invariant in a pure classification problem. Embedded in a statistical context, this means that the pdf

$$p(\mathbf{c}|\Omega_\kappa, \mathbf{a}_l) \text{ and } p(\mathbf{c}|\mathbf{a}_l) \tag{10}$$

can be estimated during training. A common approach is to make some assumption about the underlying distribution and to estimate the parameters of the distribution. For the estimation one approach is to choose some or all camera parameters \mathbf{a}_l in a supervised learning step. A feature extraction mechanism transforms the image $\mathbf{f}_{\mathbf{a}_l}$ into a feature \mathbf{c} . All that matters is that $p(\mathbf{c}|\Omega_\kappa, \mathbf{a}_l)$ and $p(\mathbf{c})$ must be represented and estimated during a training step 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 (10) have been estimated as already described in Section 2 the

mutual information can be used to decide on the optimal parameters \mathbf{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 mutual information in the notation given above is

$$I(\Omega; \mathbf{c}|\mathbf{a}_l) = \sum_{\kappa} \int_{\mathbf{c}} p_{\kappa} p(\mathbf{c}|\Omega_{\kappa}, \mathbf{a}_l) \log \frac{p(\mathbf{c}|\Omega_{\kappa}, \mathbf{a}_l)}{p(\mathbf{c}|\mathbf{a}_l)} d\mathbf{c} \quad , \quad (11)$$

with $\kappa = 1 \dots K$ being the class label. The value of $I(\Omega; \mathbf{c}|\mathbf{a}_l)$ 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 following experiments the range of the feature \mathbf{c} is discretized, so that the integration in (11) is reduced to a summation over the discrete values \mathbf{c}_i

$$I(\Omega; \mathbf{c}|\mathbf{a}_l) = \sum_{\kappa} \sum_{\mathbf{c}_i} p_{\kappa} p(\mathbf{c}_i|\Omega_{\kappa}, \mathbf{a}_l) \log \frac{p(\mathbf{c}_i|\Omega_{\kappa}, \mathbf{a}_l)}{p(\mathbf{c}_i|\mathbf{a}_l)} \quad . \quad (12)$$

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 mutual information as described in [22, 21]. In Section 4 we present another way to use continuous densities and Monte Carlo evaluation of the mutual information.

Much of this paper uses as features \mathbf{c}_i the mean of the image grayscale values, which makes it even more difficult without smart sensor data acquisition to classify objects reliably. Of course, we are aware of all the well known problems of this simple possible feature, such as its sensitivity to illumination variations. Nevertheless we claim that the main benefits of our approach can be best shown with a weak feature, where obviously smart sensor data selection is necessary. However, we will also present another approach based on eigenspace classification [14] in Section 4.3.

We discretized the range of the feature values representing the mean gray value in the image from 0 to 255 into 8 equally sized intervals. Now the discrete densities $p(\mathbf{c}_i|\Omega_\kappa, \mathbf{a}_l)$ and $p(\mathbf{c}_i|\mathbf{a}_l)$ 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 \mathbf{c}_i .

4 Extension to Differential Entropy and Mutual Information

In the previous sections we have used a discrete representation of the pdf's, which simplifies the evaluation of the mutual information. We now extend the sequential decision process to use mutual information evaluated from continuous pdf.

4.1 Differential Entropy and Mutual Information

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

$$h(x) = - \int p(x) \log(p(x)) dx \quad (13)$$

with 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 will see later on that differential version of the mutual information (that is the difference between two entropies) will be always positive.

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

$$I(x; y) = h(x) - h(x|y) = \int p(x) \int p(y|x) \log \left(\frac{p(y|x)}{p(y)} \right) dy dx \quad (14)$$

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

One practical problem with the definition of the differential mutual information is the evaluation of the double integral term. Even for Gaussian distributed random variables there exists no closed form solution for eq. (14). In the next section we will show that eq. (14) can be evaluated under very general assumptions using Monte Carlo methods. There is another way to treat continuous densities and differential entropy and mutual information by quantization of the continuous random variables. It can be shown that the discrete entropy of an n -bit quantization of a continuous random variable is approximately $h(x) + n$ with $h(x)$ being the continuous entropy [5]. For the mutual information it turns out to be even simpler to find a relation between the discrete and the differential versions since

$$I(x^\Delta; y^\Delta) = H(x^\Delta) - H(x^\Delta|y^\Delta) \quad (15)$$

$$\approx h(x) + n - (h(x|y) + n) \quad (16)$$

$$= I(x; y) \quad (17)$$

with x^Δ and y^Δ being the n -bit quantized versions of the continuous random variables x and y respectively. In other words for practical considerations one could treat differential mutual information by using a suitable quantization of the continuous pdf's and evaluating the discrete mutual information. This relationship might also serve as justification of the discretization of the feature space done in Section 3.

4.2 Monte Carlo Evaluation of Mutual Information

To avoid quantization of a continuous random variable (as was done in Section 3) we turn to the computation of mutual information by Monte Carlo sampling. Looking at eq. (14) shows an interesting fact of the mutual information that can be exploited during evaluation. Eq. (14) can be rewritten as

$$I(x; y) = E_{p(x)} \left[E_{p(y|x)} \left[\log \left(\frac{p(y|x)}{p(y)} \right) \right] \right] \quad (18)$$

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

$$\hat{E}_{p(Z)} [f(Z)] = \frac{1}{n} \sum_{z_i} f(z_i) \quad (19)$$

for $1 \leq i \leq n$. The law of large number states that $\hat{E}_{p(z)} [f(z)]$ will converge to $E_{p(z)} [f(z)]$ with probability one [19]. The estimated Monte Carlo standard error of $\hat{E}_{p(Z)} [f(Z)]$ is

$$\frac{1}{\sqrt{n}} \sqrt{\frac{\sum (f(z_i) - \hat{E}_{p(Z)} [f(Z)])^2}{n-1}} \quad (20)$$

Having in mind the relationship of eq. (18) and eq. (14) one can determine the (very general) assumptions involving the densities $p(x)$ and $p(y|x)$ that must be met for an evaluation using Monte Carlo sampling.

Proposition 1 *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, then the differential mutual information in*

eq. (14) can be approximated using eq. (18) and Monte Carlo sampling defined in eq. (19).

Proof: From the definition of the differential mutual information and the law of large numbers.

The assumptions made above 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 non-parametric representations of arbitrary densities. In [21] Parzen densities are used in the context of maximization of mutual information in an image registration framework. The maximization is performed with a stochastic gradient search called EMMA. We are less interested in a Parzen representation of arbitrary densities and more in the evaluation of the mutual information for a given continuous pdf, especially of Gaussian distributions used in the next section. However the stochastic maximization algorithm EMMA can be directly applied to our problem. This is of special interest in future work in which we plan to use a continuous representation of the actions \mathbf{a} and it becomes necessary to maximize the mutual information in a continuous parameter space. Presently we have a total number of 776 different pan/tilt/zoom positions, for which the maximum of the mutual information can be easily found by exhaustive search.

4.3 Statistical Eigenspace Classifier

In Section 6 we will show how we apply this framework of differential mutual information to view point selection for object recognition. In contrast to the Bayesian classifier based on the weak feature of mean gray value, we use in the following a more sophisticated statistical classifier that is derived from an eigenspace approach.

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). It is known that PCA minimizes the mean quadratic reconstruction error. The mapping Φ from high dimensional image space \mathbf{f} to low dimensional feature space $\mathbf{c} = \Phi \mathbf{f}$ is defined by computing the eigenvalues of the matrix $\mathbf{Q} = \mathbf{F}\mathbf{F}^T$ with \mathbf{F} containing the normalized training images of the different object from the data base. The eigenvectors φ_l that correspond to the k largest eigenvalues of \mathbf{Q} then form the matrix $\Phi = (\varphi_1, \varphi_2, \dots, \varphi_k)^T$. Instead of using one eigenspace for all object classes there exist also approaches that estimate for each object class Ω_κ a transformation matrix Φ_κ using only images \mathbf{f}_i from class Ω_κ . In the following we use one eigenspace for all classes.

After computing the transformation matrix Φ , each training image \mathbf{f}_i is projected into the eigenspace. The resulting feature vector $\mathbf{c}_i = \Phi \mathbf{f}_i$ is stored together with the class label and sometimes pose parameters of the object in image \mathbf{f}_i . During classification an image \mathbf{f} is projected into the eigenspace and the decision is made for that class (and pose) for which the stored feature vector \mathbf{c}_i has minimum distance to the vector $\Phi \mathbf{f}$. Sometimes curves are fitted to the discrete positions of the feature vectors \mathbf{c}_i for one object class to define a manifold for that class. The minimum distance is then computed to the manifold and not to the stored features.

For the selection of the best pan/tilt/zoom position of the camera defined by the maximum of mutual information we need a description of the relationships of object class and image in a probabilistic framework. This means that we need densities $p(\mathbf{c}|\Omega_\kappa)$ for each object class. Although there exists a very promising approach for probabilistic principal component analysis that results directly in the desired densities [20], for simplicity our implementation follows the approach of [2].

In the following we assume that for a given transformation Φ images \mathbf{f} from class Ω_κ are

Gaussian distributed in the feature space \mathbf{c} . In other words, one can define $p(\mathbf{c}|\Omega_\kappa)$ by

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

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 Φ .

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

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

Finally for n classes m different pan/tilt/zoom positions \mathbf{a} we end up with a total number of $m \cdot n$ Gaussian distributions, which are necessary for the computation of the differential mutual information in eq. (14). In our case $m = 776$ and $n = 9$.

5 Related Work

Information theoretic concepts have not been of particular interest in the computer vision community for a long time. Only recently these concept are recognized and applied in different applications, covering image registration [21], view point selection in object recognition [17] and feature extraction [8].

The work that is closest to our approach and that actually has been the motivation and starting point for us is the approach of active object recognition described in [17]. The authors present an active object recognition scheme based on the transformation to optimally place receptive

fields over the object of interest. The main difference to our work is that they neither perform a sequential decision process nor take the a priori probability into account. They assume that each object is equally probable. However, they perform not only classification but also localization of objects in 3D.

In the area of view point selection for object recognition two other approaches can be found, the first using Reinforcement Learning as the basis of view point selection. In [6] a reward is defined based on the difference in the distance in Eigenspace between the best and second best hypotheses. During an unsupervised training step the best sequence of view points is trained automatically. In [2] an appearance based classifier is applied together with a view point selection scheme based on the average loss in entropy. Although the authors apply a sequential fusion scheme it remains unclear how the evaluation of the average loss in entropy is done in the continuous case.

In the area of image registration the work of [21] is a good example for 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 mutual information. 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 [22].

In [8] an information theoretic approach for feature extraction is presented motivated by Fano's inequality for the error rate in classification. As in the work of [21] they represent the continuous pdf's by Parzen window densities. The work can be seen as a practical realization of a feature selection scheme based on the mutual information as 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.

In the general area of active vision and action selection information theoretic concepts have been investigated recently. Examples are active localization of robots [9], active view point selection for object recognition [1], and sensor planning for active object search [23].

The most rigorous application of information theory in image processing and computer vision can be found in [10]. The image formation process in 3D is completely embedded in an information theoretic framework. The whole process is modeled as a channel in Shannon's sense to come up with the best possible picture at the lowest data rate. The developments lead to the critical factors that limits the image formation process in a mathematical derived manner. Although computer vision problems can never be described in such a formal way, the work in [10] deals as a good example of how and when to use analytical derived densities describing the world.

None of the reviewed work takes the a priori probability of the object into account. Also, to our knowledge no approach exists that performs a sequential decision process to systematically reduce uncertainty over time. Finally, we believe that our iterative improvement of state estimation based on differential mutual information using parametric pdf's Monte Carlo sampling is new.

6 Experimental Results

In this section we summarize the experiments performed for sequential view point selection in object recognition. More experiments with real data and simulated camera parameters appear in [7]. Here we describe experiments with real camera movements using a discrete density representation and the Bayesian classifier based on the mean gray value as feature. In Section 6.2 we then present experiments using the statistical eigenspace approach as classifier, continuous densities, and Monte Carlo evaluation of the mutual information.

6.1 Parameter Selection Using Discrete Mutual Information

In this section classification results from experiments with a real pan/tilt/zoom camera are presented. In Figure 5 the data set is shown; it consists of nine different objects. Some of the objects have been modified so that they look similar. Two objects are so similar (objects 2 and 5), that a distinction using the mean gray value as feature is impossible (the central patch is actually a different color). From Figure 5 it is obvious that with this impoverished feature a classification without smart focal length and gaze control is impossible. In particular the quantized mean gray value, used as feature, is the same for all objects in the overview images shown in Figure 5 (up to our level of discretization).

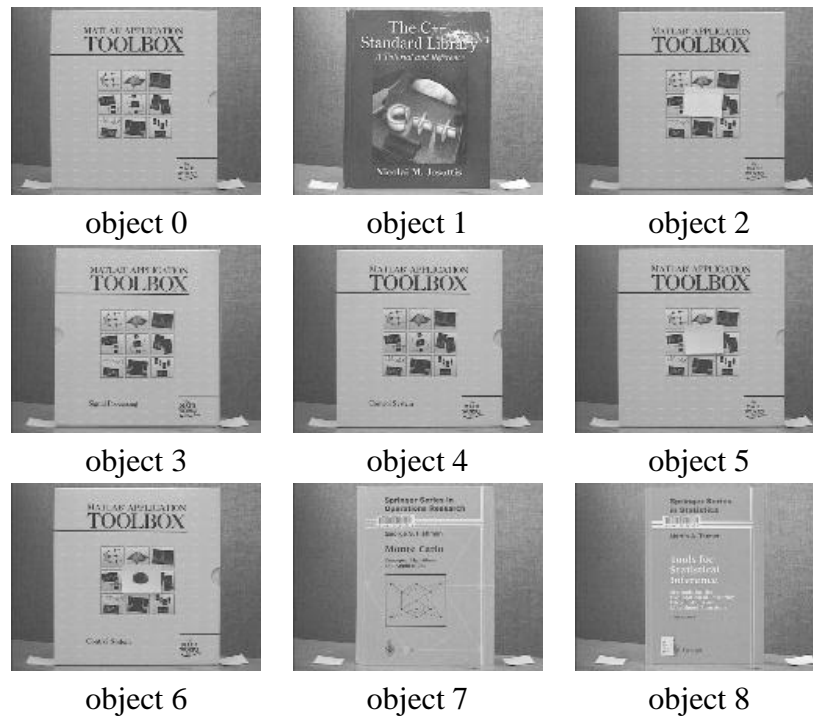


Figure 5: Data set for classification using zoom planning.

To perform classification, the following quantities from Section 3 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 $\mathbf{a} = (p, t, z)^T$, with p , t and z being the pan, tilt and zoom position 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 Figure 6. The range of pan and tilt is dependent on the selected focal length to avoid imaging the background. Again, pan and tilt position are discrete values.



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

During training, the different densities in (11) must be estimated. The most important part is the estimation of the conditional density $p(o|x, \mathbf{a})$. Thus, for all objects in a 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 position was set for each object between 100 and 10000 times), the density $p(o|x, \mathbf{a})$ can be estimated by computing the relative frequency of the observed feature o .

The experiments were performed as follows (compare also Figure 3):

1. Initialization: the distribution over the 9 classes has been initialized uniformly, to take into account that a priori (and from the overview image) no information favoring any class is available.
2. Parameter selection: based on the a priori probability the best pan/tilt position and focal length is computed using the maximum mutual information criterion (using eq. (6)).
3. Imaging and feature extraction: the pan/tilt/focal length parameters are set for the camera. An image is taken and the feature (quantized mean gray value) is extracted.
4. Bayes decision: Bayes formula is used to compute the a posteriori probability for the 9 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.

In Figure 6 the range in focal length during the experiments is visualized. Between the close up view and overview view the focal range has been discretized uniformly into 6 different positions. The pan and tilt positions have also been discretized, depending on the chosen focal length, i.e. at least 50% of the image contained the object itself. Thus, the density $p(o|x, \mathbf{a})$ is represented as a 5-dimensional table. 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.

Figure 7 depicts a typical experiment. Several more can be found in [7]. Besides the change in belief state for the 9 classes one can also see the change in entropy of the distribution over

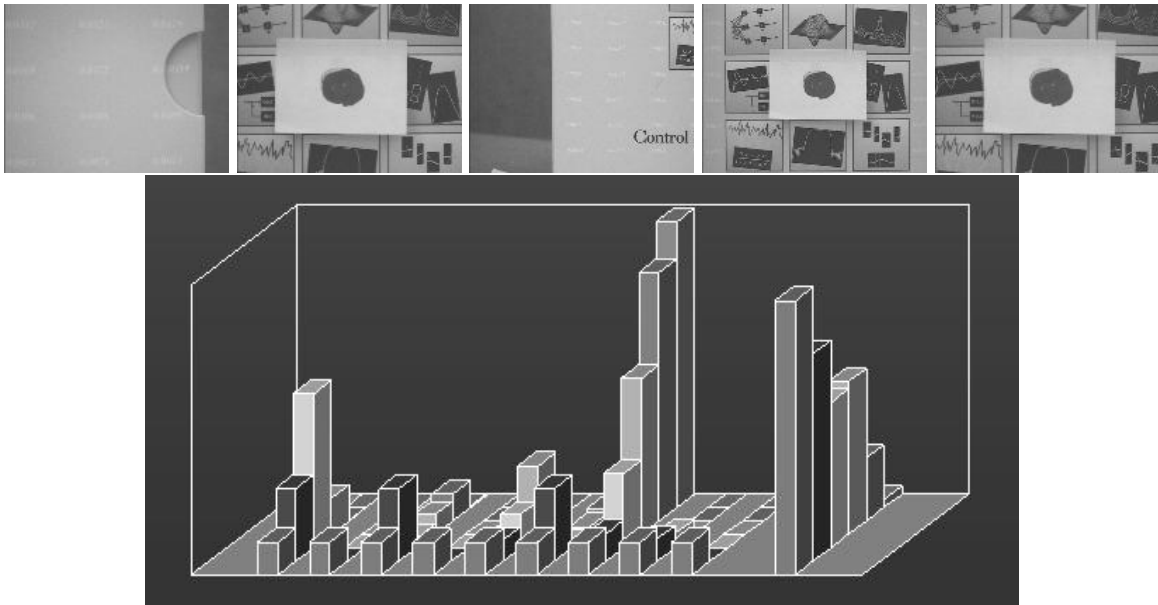


Figure 7: Object recognition using gaze control for object number 6 (the initial overview has been left out). Bottom: the change in belief state over time (from left to right, object number 0 to 8) and the change in entropy in the belief state (farthest right bars). The z-axis indicates the time step in the sequential decision process.

the classes (farthest right bars). Except for one view 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.

Figure 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 most information at the next time step. And this part is focused on during the next time interval, as can be seen in Figure 7, top row, second image. The reason for the repeated, identical look to the center (view 2 and view 5) can be explained by an mismatch between the

learned and the true underlying model for the objects. As one can see, the entropy after selected view 2 increases. Also, the maximum a posteriori probability would return object 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 return this result again the look to the center is necessary.

With higher noise in the camera parameter control, the result is that [7] the entropy never increases, but the decrease in uncertainty is dramatically reduced and also the final decision is not as unambiguous as for the experiment shown in Figure 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 for selected objects the results for random gaze and focal length control. Object 2 and 5 could not be distinguished based on the mean gray

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			

Table 1: Recognition results (in percent). First row, a low noise assumption; second row, a high noise assumption; third row (for selected objects) a random strategy.

value (compare also Figure 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 (object 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

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			

Table 2: Average number of views until decision. First row, low noise assumption; second row, high noise assumption; third row, (for selected objects) a random strategy.

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% 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%.

6.2 Parameter Selection Using Differential Mutual Information

object	rec. rate	mean no. views	mean max. prob.
o0	99.5	2.4	0.96
o1	100.0	1.0	1.00
o2	100.0	4.0	0.95
o3	100.0	2.3	0.96
o4	100.0	4.0	0.95
o5	99.2	3.5	0.97
o6	99.6	2.8	0.96
o7	100.0	1.7	0.98
o8	100.0	1.1	1.00
average	99.8	2.5	0.97

Table 3: Results for view point planning (1000 trials per object): Recognition rate, mean number of views, and maximum a posteriori probability for the right class when the decision has been made.

In this section we present experiments with the statistical eigenspace classifier and differential mutual information. As before we used the data set of the nine books shown in Figure 5. In

the training step for each pan/tilt/zoom position \mathbf{a} we took views from each object class Ω_κ to compute the transformation matrix $\Phi_{\mathbf{a}}$. Afterwards we created synthetically 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 shift in x and y position of a window centered at the image as well as pixelwise 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 \mathbf{c}_i are used for a maximum likelihood estimation of the parameters of the Gaussian densities.

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 mutual information for selection of the best next pan/tilt/zoom position \mathbf{a} , and that a statistical Eigenspace approach for classification is applied.

object	rec. rate	mean no. views	mean max. prob.
o0	83.4	9.9	0.61
o1	99.6	1.2	1.00
o2	62.4	9.8	0.65
o3	76.0	9.8	0.64
o4	66.6	10.0	0.56
o5	68.2	9.9	0.57
o6	76.7	9.7	0.63
o7	100.0	2.5	0.97
o8	100.0	2.4	0.97
average	81.4	7.2	0.73

Table 4: Results for random view point selection (1000 trials per object): Recognition rate, mean number of views, and maximum a posteriori probability for the right class when the decision has been made.

Table 3 gives the results for the view point planning strategy based on the maximum of mutual information. The decision for the next view is made by Monte Carlo evaluation (with 1000 samples) of the mutual information as described in Section 4. 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% or close to it.

It is also natural that object o1 is recognizable in any case in the first view. It is also interesting to look at the maximum a posteriori probability that one get 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 4 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 11 trials is made before the decision is forced.

Although the recognition rate for the “easy” objects — o1, o7, and o8 — is comparable to the results using view point planning, one can see that the mean number of views that are necessary to return an a posteriori probability of more the 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 reduced from 99.8% for the view point planning to 81.4% in the case of a random strategy. This result shows without any doubt that the view point selection strategy based on maximum mutual information works in practice for a standard state of the art classification method and outperforms a random strategy.

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. Despite the proposed paradigm of active vision it remains an unsolved problem in general which sensor data should be selected at a certain stage of state estimation.

Our approach tackles the problem at a different level. Instead of optimizing an estimator-specific metric (building a better edge-finder or classification algorithm) we try to reduce the uncertainty in the 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. Additionally, related to classical approaches for sequential decision making, like Markov decision processes or reinforcement learning, in our approach the time and memory consuming dynamic programming is avoided.

To measure the uncertainty in the state estimation process we have introduced a formalism based on Shannon's information theory. The goal is to reduce the uncertainty and ambiguity (variance and number of modes of the pdf) in the probability of the state over time. The optimal sequence of chosen actions would transform a uniform distribution over the state space (in the beginning of the state estimation process) to a unimodal distribution with minimum variance, whose maximum is the true state. A unimodal distribution is the best case for a wide range of state estimators, for example the Kalman filter where the underlying assumption is a unimodal (Gaussian)

distribution for the distribution over the state space.

The important quantity in our work is the conditional mutual information, conditioned on the selected camera parameters. The mutual information 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 mutual information with respect to the camera parameters returns the best action in terms of reduction in uncertainty.

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. The actions are the selection of pan/tilt and focal length of an active camera device. In contrast to related work in this area we explicitly take into account the a priori probability for the computation of the mutual information. The a priori probability at a certain time step in the state estimation process is the a posteriori probability of the previous time step. This practice relates the state estimator specific behavior to our general framework of action selection, since actions are avoided that result in observations that are not suited for an improvement in state estimation for a particular state estimator. Also the convergence of the resulting sequential decision process can be proven.

For object recognition we used a rudimentary state estimator based on the mean gray value in the captured image and discrete densities, as well as a more sophisticated classifier based on statistical eigenspace and continuous densities. The simple recognizer inherently has serious problems in distinguishing similar looking objects so that the need of smart sensor data selection becomes more obvious. Our test set consists of nine objects, with six of them looking very similar. In the experiments we have shown that our approach was able to achieve a recognition rate of more than 77% despite the weak feature chosen and the very difficult data set. Without active sensor data selection the objects could not be classified at all. Also, our approach outperforms a random strat-

egy for action selection in both the number of views necessary for classification as well as in the recognition rate. Quite similar results have been achieved in the case of the statistical eigenspace classifier. The camera parameter selection strategy based on the differential mutual information (recognition rate: 99.8%) again outperforms the random strategy (recognition rate: 81.4%). The higher overall recognition rate is due to the better features extracted in the eigenspace approach.

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 applied state estimator. 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. And last but not least, the conditional pdf's (conditioned on the action) for the observations and the likelihood function must be known or estimated in a training step. As it can be verified easily the 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. This means that assumptions for the underlying distributions must be made and verified. The estimation of the parameters of the densities is not a trivial problem, especially in higher dimensional spaces (state, feature, and action). Secondly, since we do not optimize or adapt the parameters of the state estimator the sequential decision process will not improve state estimation if the state estimator systematically returns wrong or strongly biased state estimates. The criterion — reducing uncertainty and ambiguity — will be still optimized, although the result of state estimation is not improved. 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 [22, 21] an approach, EMMA, has been developed for maximizing the mutual information of two random variables represented by a Parzen window density for alignment of images of different modalities. Such an algorithm for maximization of the mutual information becomes important when we extend the discrete actions space to a continuous one. Finally, we are working on extending the presented framework to state estimation in dynamic systems.

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. Active object recognition in parametric eigenspace. In *British Machine Vision Conference 1998*, volume 2, pages 629–638, 1998.
- [3] H. Borotschnig, L. Paletta, M. Prantl, and A. Pinz. A comparison of probabilistic, possibilistic and evidence theoretic fusion schemes for active object recognition. *Computing*, 62:293–319, 1999.
- [4] D.A. Cohn, A. Ghahramani, and M.I. Jordan. Active learning with statistical models. *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- [5] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications. John Wiley and Sons, New York, 1991.
- [6] F. Deinzer, J. Denzler, and H. Niemann. Viewpoint Selection - A Classifier Independent Learning Approach. In *IEEE Southwest Symposium on Image Analysis and Interpretation*, 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 Department, University of Rochester, 2000.
- [8] J. Fisher and J.C. Principe. A nonparametric method for information theoretic feature extraction. In *DARPA Image Understanding Workshop*, New Orleans, 1997.
- [9] D. Fox, W. Burgard, and S. Thrun. Active markov localization for mobile robots. Technical report, Carnegie Mellon University, 1998.

- [10] F.O. Huck, C.L. Fales, and Z. Rahman. An information theory of visual communication. *Philosophical Transactions: Mathematical, Physical and Engineering Sciences*, (354):2193–2248, 1996.
- [11] M. Isard and A. Blake. Contour tracking by stochastic propagation of conditional density. In A. Blake, editor, *Computer Vision - ECCV 96*, pages 343–356, Berlin, Heidelberg, New York, London, 1996. Lecture Notes in Computer Science.
- [12] L.P. Kaelbling, M.L. Littman, and A.R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101(1–2):99–134, 1998.
- [13] R.E. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, pages 35–44, 1960.
- [14] H. Murase and S. Nayar. Visual Learning and Recognition of 3–D Objects from Appearance. *International Journal of Computer Vision*, 14:5–24, 1995.
- [15] H. Niemann. *Pattern Analysis and Understanding*, volume 4 of *Series in Information Sciences*. Springer, Berlin Heidelberg, 1990.
- [16] S.J. Russel and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, 1994.
- [17] B. Schiele and J.L. Crowley. Transinformation for active object recognition. In *Proceedings of the Sixth International Conference on Computer Vision*, Bombay, India, 1998.
- [18] R.S. Sutton and A.G. Barto. *Reinforcement Learning*. A Bradford Book, Cambridge, London, 1998.
- [19] M.A. Tanner. *Tools for Statistical Inference*. Springer Verlag, London, Berlin, Heidelberg, New York, Paris, Tokyo, Hong Kong Budapest, 1993.
- [20] M.E. Tipping and C.M. Bishop. Mixtures of probabilistic principal component analysers. *Neural Computation*, page to appear, 2000.
- [21] P. Viola and W.M. Wells III. Alignment by maximization of mutual information. *International Journal of Computer Vision*, 24(2):137–154, 1997.
- [22] P.A. Viola. Alignment by maximization of mutual information. Technical Report AI Technical Report No. 1548, MIT Artificial Intelligence Laboratory, 1995.
- [23] Y. Ye. Sensor planning for object search. Technical Report PhD Thesis, Department of Computer Science, University of Toronto, 1997.