

# Optimal Selection of Camera Parameters for State Estimation of Static Systems: An Information Theoretic Approach

J. Denzler                      C. Brown

The University of Rochester  
Computer Science Department  
Rochester, New York 14627

Technical Report 732

28th August 2000

## 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 consider a framework based on Shannon's information theory and select the camera parameters that maximize the mutual information, i.e. optimize the information that the captured image conveys about the true state of the system. The technique implicitly 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 using an object recognition scenario and an active pan/tilt/zoom camera. During the sequential decision process the camera looks to parts of the object that allow the most reliable distinction of similar looking objects. We performed experiments with discrete density representation as well as with continuous densities and Monte Carlo evaluation of the mutual information. The results show that 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.

---

This work has been funded by the German Science Foundation (DFG) under grant DE 735/1, and was partially supported by NSF grant EIA-9972881 and CAT/NYSSTF grant EEC-9813002.

# 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 [Kalman, 1960; Cohn *et al.*, 1996] 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 if defined in terms of reduction of uncertainty and ambiguity. A formal description of this kind of optimality is presented in Section 2.

We formulate state estimation in a probabilistic framework. Instead of estimating a “true state” (object class, object pose, dynamics of an object) we look at the state as a probability density function (pdf’s sometimes called belief state [Russel and Norvig, 1994]) over the state space. We use the maximum a posteriori (MAP) decision method — a common approach for returning the most likely state given the a priori information and the information from the sensors.

Uncertainty in state estimation will in general increase the variance, while ambiguity increases the number of modes of the belief state. 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 uniform prior distribution  $p(\mathbf{x}_t)$  over the state space  $\mathbf{x} \in \mathbb{R}^n$  (i.e. 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.

In the following we look at one special class of camera actions  $\mathbf{a}$ , 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

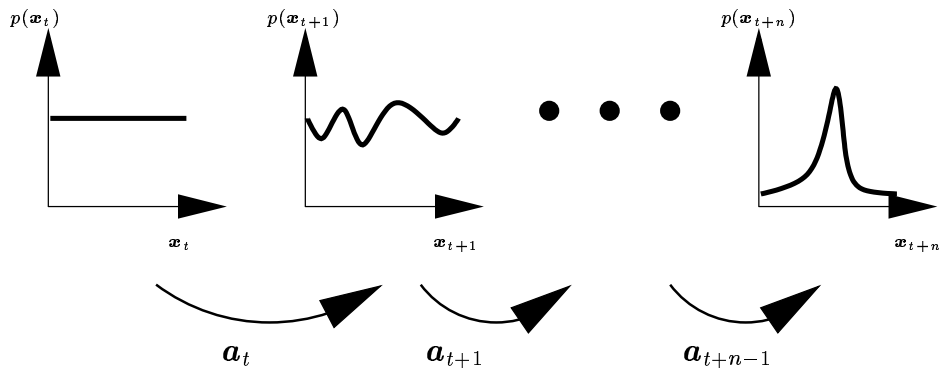


Figure 1: General principle: reduce uncertainty and ambiguity (variance and multiple modes) in the pdf of the state  $\mathbf{x}_t$  by choosing appropriate actions  $\mathbf{a}_t$ , or to be more precise, by selecting the right sensor data.

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 ideas presented in this paper are not limited to static state estimation. We are currently working on an extension to the case of state estimation in dynamic systems.

In our approach the usefulness of a chosen focal length setting of one camera is predicted by computing the mutual information or transinformation between the state and the observation distribution and maximizing this quantity. From an information theoretic point of view the maximization of the mutual information means that the data (in this case influenced by the focal length and pan/tilt selection) is chosen in a way that contributes most to reducing the uncertainty in the estimation of the unknown state.

The basis for the sensor data selection is the current information about the state of the object, for example, the class of a certain object seen in the image. In that case, the system estimates the most valuable focal length, which is used to take a new image. The new image is classified and as a result one gets a new a posteriori probability distribution over the object classes. This new probability can be treated as a priori information for the next time step, yielding a sequential decision process. The sequential decision process is optimal in the sense that it uses mutual information to make optimal parameter decisions at each step. The decisions in turn reduce uncertainty and ambiguity over time, and converge.

The key point in this work is the integration of sensor models and the effect of actions into a probabilistic framework. In the case of object recognition the effect of actions is learned in advance. In terms of information theory and the source-channel-destination framework (Figure 3) changing the sensor parameters then means changing the properties of the transmission medium, i.e. the channel. Thus, the optimal selection of sensor data can be integrated in a natural way into the framework of information theory. Currently no external costs, like time, the cost of camera movements or computational costs if using higher resolution are taken into account. Nevertheless there is no restriction in the general framework that would prevent us from integrating such costs.

Finally, as already mentioned this approach is not limited to pan, tilt, and zoom adjustments. The

approach can be extended to several other possible actions, 3-D movements of a camera, control of lighting or iris control, and tuning of the focus of a camera. It is also independent of the underlying decision algorithm (e.g. the underlying classification algorithm). As a matter of fact the approach is able to make use of any algorithm-specific properties that depend on the actively selected parameter. To give an example consider the case of template matching for pattern recognition. Unless the size of the template can be scaled with respect to the size of the object in the image, there will be only a small range of object sizes for which the recognition will work well. The size of the object in the image depends on the chosen focal length. Thus, the active selection of the camera parameter is also influenced by the algorithm specific demands: in the example, the size of the template.

The paper is structured as follows. In Section 2 a formal statement of the problem is given (appendix A gives a short introduction to the most important terms from information theory). 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

Most problems in computer vision, especially dynamic problems, cycle (either explicitly or implicitly) through a state estimation and action selection stage (see Figure 2). 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. An example for the latter uncertainty is the selection of a certain focal length. Due to mechanical inaccuracies, especially for low-cost off-the-shelf cameras, it is very unlikely that the selected position will be reached exactly. 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. To give an example, a common assumption for the noise processes is Gaussian noise.

The probabilistic framework also allows us to use probability distribution functions 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 the recent years the belief state representation has been extensively used and studied in the context of particle filters [Isard and Blake, 1996].

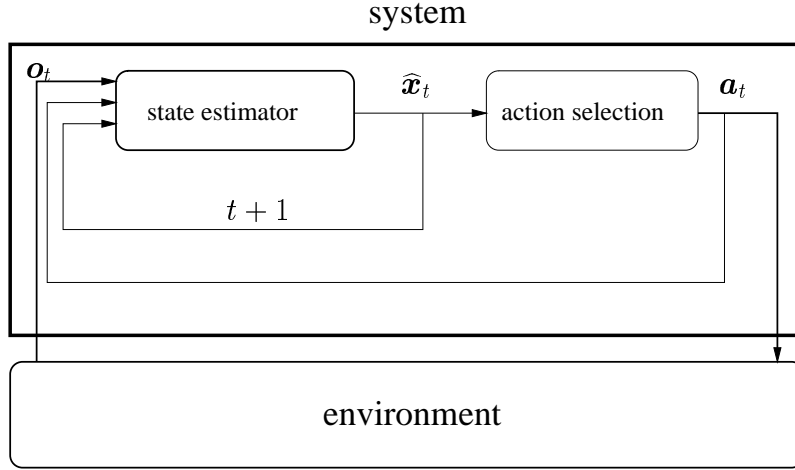


Figure 2: General loop of state estimation and action selection.

Two examples can be given to clarify our scenario. In object tracking the 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. Object tracking is also an example of a time varying state in a dynamic system. In object recognition the 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 [Deinzer *et al.*, 2000; Borotschnig *et al.*, 1999; Schiele and Crowley, 1998]. This second example can be treated as an instance of a static state. Thus the problem of optimal sensor data selection occurs in static as well as in dynamic systems; we will restrict ourselves to the problem of state estimation in static systems. An example of a static system will be discussed in Section 3, where the camera position together with the focal length are the parameters that must be optimized.

Figure 3 gives the main elements of our approach. It shows the transmission of a state  $\mathbf{x}$  over a channel. At the other end of the channel an observation  $\mathbf{o}$  is made. The system gets as input an 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 3 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 density function is abbreviated with  $p(\mathbf{x}_t)$  in

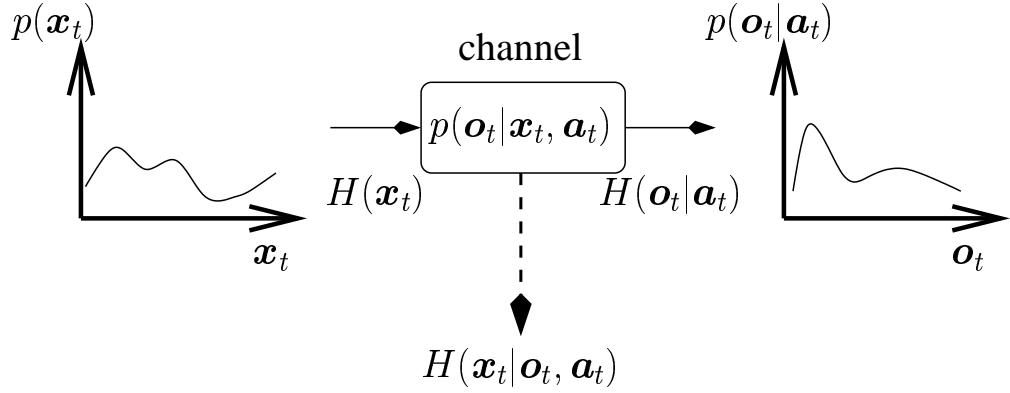


Figure 3: Input and output relation in the channel model and some of the important entropies  $H$  describing the information content. The state estimator (compare Figure 2) that estimates the belief state  $\mathbf{x}_t$  based on the observation  $\mathbf{o}_t$  is missing in this figure.

Figure 3. With that probability density function 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 (Appendices A and B define other important quantities related to information theory). The entropy measures the amount of uncertainty in a random experiment using the probability density function  $p(\mathbf{x}_t)$ . 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 theory 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 probability density function  $p(\mathbf{o}_t | \mathbf{a}_t)$  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 the appendices A and B. 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

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

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 the proposed 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. This has been already motivated in Figure 1 in Section 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. One example of a rigorous development of a statistical model for an image formation process can be found in [Huck *et al.*, 1996]. This work may serve as an example for when to construct and use an analytical statistical model in contrast to a learned one.

In the next section we will show the capability of our approach using a classical problem in computer vision, namely object recognition. We will show how optimal sensor parameters can be chosen automatically. The reasons for the choice of this problem are twofold. First, we can show how the general framework of using mutual information for action selection can be mapped to improving state estimation in static systems by smart sensor parameter selection. Second, the problem of object recognition is an application where the use of camera parameter selection can be easily motivated and can be used immediately to improve classification results.



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

### 3 Object Recognition

We are interested in the optimal selection of sensor data for object recognition. During the past, most object recognition systems made their decision based on a single image. A problem that arises from a decision based on a single image is that ambiguities between objects cannot always be resolved. One example is the two cups shown in Figure 4. In the first view, 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 of course is how can we identify such views automatically. One straightforward but not practical approach would be to take all possible views or a set of random views of the object. The hope would be that this image set contains enough images that are simple in the sense of correct recognition. A better approach of course 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 (it is worth noting that having the best view still does not necessarily result in a correct classification).

In the literature there exists some work on active object recognition. For example [Schiele and Crowley, 1998] used essentially the same criterion that we introduced in Section 2. The major difference to the approach developed here lies in the usage of the a priori information. In [Schiele and Crowley, 1998] this probability density function is assumed to be uniformly distributed. In our approach we explicitly take the a priori probability density function into account by performing a sequential decision process, where the a priori information changes over time. The goal is finally to get a unimodal a posteriori distribution, with the maximum being the true class and with a minimal variance. A knowledge based approach for 3D object recognition using Bayesian networks has been presented in [Krebs *et al.*, 1998]. In [Deinzer *et al.*, 2000] an approach based on reinforcement learning for viewpoint selection is presented. Based on a reward definition, which measures how distinguishable the objects are, given a certain view, viewpoint planning is done by maximizing the reward. The mapping from the state to the chosen action is trained automatically

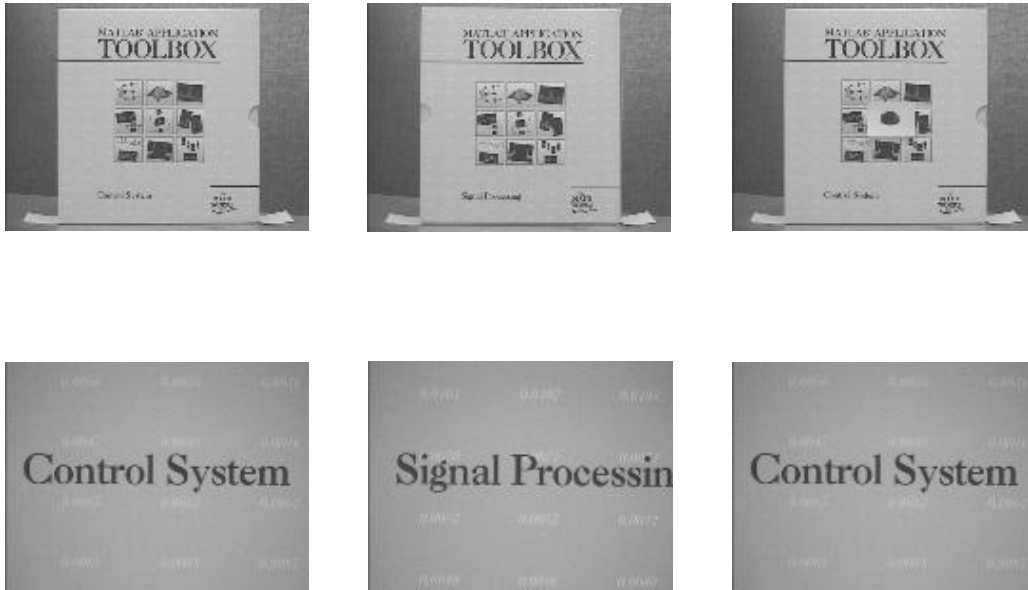


Figure 5: Three different objects, where the trade off between global overview and local, detailed inspection can be shown. First row: global overview images. Second row: detailed inspection of the lower left part of the objects. For the recognition algorithm we use it is possible to detect differences between objects (1,2) and (3) easily.

using an eigenspace approach for the classification step [Murase and Nayar, 1995] and Monte Carlo reinforcement learning [Sutton and Barto, 1998] for approximating the reward function.

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. An example is shown in Figure 5. The first row shows the global overview images, where immediately the difference between object number three and the first two objects can be seen. More difficult is the distinction between objects one and two. The distinction can only be done by zooming toward the writing on the lower left part of the boxes. Unfortunately, the same writing appears on object numbered one and three. As a consequence, zooming does not become useful until object three has been ruled out.

### 3.1 Mutual Information for Camera Parameter Selection in Object Recognition

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  $\alpha_l$  shall summarize the focal length as well as 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 zero position is defined as that pan/tilt position where the optical axis hits the center of the object. It is easy to find automatically, at least for single object scenes.

During a training step for each camera parameter  $\mathbf{a}_l$  we observe for each object  $\Omega_\kappa$ ,  $\kappa = 1 \dots K$ , a certain feature  $\mathbf{c}$ . The class label  $\Omega_\kappa$  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 following probability density functions

$$p(\mathbf{c}|\Omega_\kappa, \mathbf{a}_l) \quad (7)$$

and

$$p(\mathbf{c}|\mathbf{a}_l) \quad (8)$$

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  $\mathbf{c}$ . In the following we use as feature the mean value of the gray values in the image. Any other more sophisticated feature can be used instead. 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. We use this simple, weak, scalar feature because it is easy to extract and learn, and because it illustrates that even such a simple feature is effective if the camera parameters are chosen using our scheme.

As soon as the densities in (7) and (8) 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  $\Omega_\kappa$ . The new camera parameters are used to take a new image. The mutual information in the notation given above is

$$I(\Omega; \mathbf{c}|\mathbf{a}_l) = \sum_{\kappa} \int_{\mathbf{c}} p_{\kappa} p(\mathbf{c}|\Omega_{\kappa}, \mathbf{a}_l) \log \frac{p(\mathbf{c}|\Omega_{\kappa}, \mathbf{a}_l)}{p(\mathbf{c}|\mathbf{a}_l)} d\mathbf{c} \quad , \quad (9)$$

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. The proof is simple. Since the mutual information can be written as

$$I(\Omega; \mathbf{c}) = H(\Omega) - H(\Omega|\mathbf{c}) \quad (10)$$

the maximum value is reached (assuming constant  $p_{\kappa}$ ) if  $H(\Omega|\mathbf{c})$  is zero (i.e. the uncertainty about the class  $\Omega$  is zero, if the feature  $\mathbf{c}$  is observed) and therefore the maximum is  $H(\Omega) = -\sum p_{\kappa} \log p_{\kappa}$  by definition.

For the following experiments the range of the feature  $\mathbf{c}$  is discretized, so that the integration in (9) is reduced to a summation over the discrete values  $\mathbf{c}_i$

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

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 [Viola, 1995; Viola and Wells III, 1997]. In

Section 4 we present another way to use continuous densities and Monte Carlo evaluation of the mutual information.

For the features  $\mathbf{c}_i$  we take the mean of the gray values within the image, 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 [Murase and Nayar, 1995] 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  $\Omega_\kappa$  and  $\mathbf{c}_i$ . The evaluation of the mutual information in (9) becomes now — having  $p(\mathbf{c}_i|\Omega_\kappa, \mathbf{a}_l)$  and  $p(\mathbf{c}_i|\mathbf{a}_l)$  — a linear function in  $p_\kappa$ , the a priori probabilities for the object classes. In other words, the a priori probability has the most influence on the decision of the next camera parameter selection, since the  $p_\kappa$  are the only free parameters. Actually, such a dependency is wanted.

### 3.2 Sequential Decision Making

At the beginning of classification we assume that no a priori information is available supporting a certain class. Thus, the  $p_\kappa$  are initialized uniformly. If reliable, non-uniform priors are known, of course we could use them at this point. The evaluation of the mutual information returns in this case the camera parameters that allow the best distinction of the classes based on the training information that we call the optimal camera parameters. The resulting image is used to classify the object using Bayes rule, which computes for each class  $\Omega_\kappa$  the a posteriori probability

$$p(\Omega_\kappa|\mathbf{c}, \mathbf{a}_l) = \frac{p(\mathbf{c}|\Omega_\kappa, \mathbf{a}_l)p_\kappa}{p(\mathbf{c}|\mathbf{a}_l)} \quad (12)$$

and which decides for the class with maximum a posteriori probability. The vector of the a posteriori probabilities is now taken as a priori probabilities  $p'_\kappa = p(\Omega_\kappa|\mathbf{c}, \mathbf{a}_l)$  for the next time step, in which new camera parameters will be selected. Thus, 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 6. Since each step is optimal in the sense of the mutual information the whole process is optimal, too.

Looking at the equation for the mutual information, one can see that the mutual information can be written as a linear function

$$I(\Omega; \mathbf{c}|\mathbf{a}) = \sum_{\kappa}^K e_\kappa(\mathbf{a})p_\kappa = \mathbf{e}(\mathbf{a})\mathbf{p} \quad (13)$$

with

$$\mathbf{p} = (p_1, p_2, \dots, p_K)^T \text{ and } \mathbf{e}(\mathbf{a}) = (e_1(\mathbf{a}), e_2(\mathbf{a}), \dots, e_K(\mathbf{a}))$$

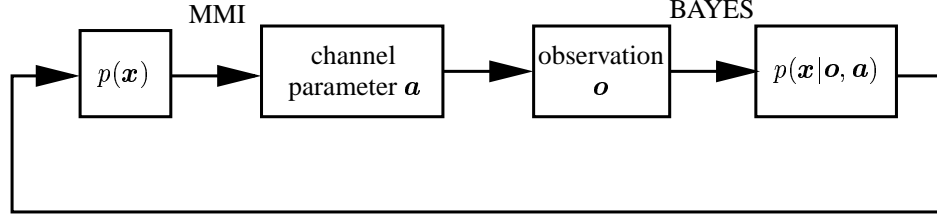


Figure 6: Sequential decision process of maximum mutual information (MMI) for camera parameter selection and Bayesian classification based on the observed feature.

where

$$e_j(\mathbf{a}) = \sum_{\mathbf{c}_i} p(\mathbf{c}_i|\Omega_j, \mathbf{a}) \log \frac{p(\mathbf{c}_i|\Omega_j, \mathbf{a})}{p(\mathbf{c}_i|\mathbf{a})} . \quad (14)$$

This measure depends during recognition on the a priori probability  $\mathbf{p}$  since the other conditional densities  $p(\mathbf{c}_i|\Omega_j, \mathbf{a})$  are constant after training.

At the beginning of the sequential decision process (let's say at time  $t = 0$ ) the a priori probability for a certain object class  $\Omega_\kappa$ ,  $1 \leq \kappa \leq K$  is set to  $p_\kappa^{(t=0)} = p(\Omega_\kappa)^{(t=0)} = \frac{1}{K}$ , and  $\mathbf{p}^{(0)} = (p_1^{(t=0)}, \dots, p_K^{(t=0)})^T$ . The first view  $\mathbf{a}_0$  is selected based on the maximization of equation (13). From the resulting image (using camera parameter  $\mathbf{a}_0$ ) the feature  $\mathbf{c}_0$  is extracted and the classification using Bayes rule returns the following a posteriori probability

$$p_\kappa^{(1)} = p(\Omega_\kappa|\mathbf{c}_0, \mathbf{a}_0) = \frac{p(\mathbf{c}_0|\Omega_\kappa, \mathbf{a}_0)p(\Omega_\kappa|\mathbf{a}_0)^{(0)}}{p(\mathbf{c}_0|\mathbf{a}_0)} = \quad (15)$$

$$= \frac{p(\mathbf{c}_0|\Omega_\kappa, \mathbf{a}_0)p(\Omega_\kappa)^{(t=0)}}{p(\mathbf{c}_0|\mathbf{a}_0)} \quad (16)$$

where  $\mathbf{a}_0$  is given by

$$\mathbf{a}_0 = \underset{\mathbf{a}}{\operatorname{argmax}} I_0(\Omega; \mathbf{c}|\mathbf{a}) . \quad (17)$$

The step from equation (15) to (16) is justified by the assumption 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. Since more information is available about the object one can interpret this also as having ruled out some of the possible objects. The mutual information in equation (13) will be changed after the first classification to

$$I_1(\Omega; \mathbf{c}|\mathbf{a}) = \sum_{\kappa}^K e_\kappa(\mathbf{a}) p_\kappa^{(1)} = \mathbf{e}(\mathbf{a}) \mathbf{p}^{(1)} \quad (18)$$

with

$$\mathbf{p}^{(1)} = \left( \frac{p(\mathbf{c}_0|\Omega_1, \mathbf{a}_0)p(\Omega_1)^{(0)}}{p(\mathbf{c}_0|\mathbf{a}_0)}, \frac{p(\mathbf{c}_0|\Omega_2, \mathbf{a}_0)p(\Omega_2)^{(0)}}{p(\mathbf{c}_0|\mathbf{a}_0)}, \dots, \frac{p(\mathbf{c}_0|\Omega_K, \mathbf{a}_0)p(\Omega_K)^{(0)}}{p(\mathbf{c}_0|\mathbf{a}_0)} \right)^T \quad (19)$$

and  $\mathbf{e}(\mathbf{a})$  as before. In general after the  $n$ th view planning step one gets

$$I_n(\Omega; \mathbf{c}|\mathbf{a}) = \sum_{\kappa}^K e_{\kappa}(\mathbf{a}) p_{\kappa}^{(n)} = \mathbf{e}(\mathbf{a}) \mathbf{p}^{(n)} \quad (20)$$

with

$$\mathbf{p}^{(n)} = \left( \begin{array}{c} \frac{p(\mathbf{c}_n|\Omega_1, \mathbf{a}_n)p(\Omega_1|\mathbf{a}_{n-1}, \dots, \mathbf{a}_0)}{p(\mathbf{c}_n|\mathbf{a}_n)}, \\ \frac{p(\mathbf{c}_n|\Omega_2, \mathbf{a}_n)p(\Omega_2|\mathbf{a}_{n-1}, \dots, \mathbf{a}_0)}{p(\mathbf{c}_n|\mathbf{a}_n)}, \\ \vdots \\ \frac{p(\mathbf{c}_n|\Omega_K, \mathbf{a}_n)p(\Omega_K|\mathbf{a}_{n-1}, \dots, \mathbf{a}_0)}{p(\mathbf{c}_n|\mathbf{a}_n)} \end{array} \right)^T \quad (21)$$

and

$$\mathbf{a}_{n-1} = \operatorname{argmax}_{\mathbf{a}} I_{n-1}(\Omega; \mathbf{c}|\mathbf{a}) \quad . \quad (22)$$

Here, the plausible assumption is made that the distribution of the features of view  $n$  depend only on the class and the chosen view, but not on the past views. Using absolute instead of relative camera parameters the assumption is certainly valid. Equations (20), (21) and (22) define the recursive viewpoint selection.

A classical way to describe a sequential decision process known from literature is the so called 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 a brilliant textbook on reinforcement learning by Sutton [Sutton and Barto, 1998]). Recently, also the partially observable case [Kaelbling *et al.*, 1998] has been treated but still by either applying dynamic programming or directly solving the so called Bellman equations. In contrast to the Markov decision process approach, in our work the time and memory intensive dynamic programming is avoided. However in our approach, estimation of the necessary statistical information (eq. (11)), is not a trivial task. This estimation might be unnecessary if such knowledge is implicitly provided by the state estimator.

### 3.3 Convergence and Optimality of the Sequential Decision Making

The experiments in the next section show that the sequential decision making process will converge in practice. Actually this convergence can also be formally proved. The proof is given in appendix C. One consequence of the proof is that under certain assumptions that are difficult to verify the sequential decision process is guaranteed to return the right class. 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. 10). 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.

## 4 Extension to Differential Entropy and Mutual Information

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

### 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 [Cover and Thomas, 1991]

$$h(x) = - \int p(x) \log(p(x)) dx \quad (23)$$

with the integral being evaluated over the support set of the random variable  $x$ . For a uniform distribution  $p(x) = \frac{1}{a}$  for  $x \in [0; a]$ , the entropy

$$h_1(x) = \log(a) \quad (24)$$

and for a Gaussian distribution  $p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{x^2}{2\sigma^2})$  the entropy

$$h_2(x) = \frac{1}{2} \log(2\pi e\sigma^2) \quad (25)$$

can be computed in a straight forward way.

One main difference between discrete and differential entropies is that the differential entropy can become negative. This can be easily verified for the uniform distribution by setting  $a < 1$ . 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

$$I(x; y) = h(x) - h(x|y) = \int p(x) \int p(y|x) \log \left( \frac{p(y|x)}{p(y)} \right) dy dx \quad (26)$$

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. (26). In the next section we will show that eq. (26) 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 [Cover and Thomas, 1991]. For the mutual information it turns out to be even simpler to find a relation between the discrete and the differential versions since

$$I(x^\Delta; y^\Delta) = H(x^\Delta) - H(x^\Delta|y^\Delta) \quad (27)$$

$$\approx h(x) + n - (h(x|y) + n) \quad (28)$$

$$= I(x; y) \quad (29)$$

with  $x^\Delta$  and  $y^\Delta$  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 probability density functions 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. (26) shows an interesting fact of the mutual information that can be exploited during evaluation. Eq. (26) can be rewritten as

$$I(x; y) = E_{p(x)} \left[ E_{p(y|x)} \left[ \log \left( \frac{p(y|x)}{p(y)} \right) \right] \right] \quad (30)$$

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

$$\hat{E}_{p(Z)} [f(Z)] = \frac{1}{n} \sum_{z_i} f(z_i) \quad (31)$$

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 [Tanner, 1993]. The estimated Monte Carlo standard error of  $\hat{E}_{p(Z)} [f(Z)]$  is

$$\frac{1}{\sqrt{n}} \sqrt{\frac{\sum (f(z_i) - \hat{E}_{p(Z)} [f(Z)])^2}{n-1}}. \quad (32)$$

Having in mind the relationship of eq. (30) and eq. (26) 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. (26) can be approximated using eq. (30) and Monte Carlo sampling defined in eq. (31).*

**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 [Viola and Wells III, 1997] 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 probability density function, 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 previous Bayesian classifier based on the mean gray value we use in the following a more sophisticated statistical classifier that is derived from an eigenspace approach.

The idea in the following is to apply a more general classifier than the previous Bayes classifier using the mean gray value (Section 3). One well accepted method is the eigenspace approach first introduced in [Murase and Nayar, 1995]. 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  $\Phi$

$$\mathbf{c} = \Phi \mathbf{f} \quad (33)$$

from high dimensional image space  $\mathbf{f}$  to low dimensional feature space  $\mathbf{c}$  is defined by computing the eigenvalues of the matrix  $\mathbf{Q}$

$$\mathbf{Q} = \mathbf{F} \mathbf{F}^T \quad (34)$$

$$\mathbf{F} = (\mathbf{f}_0 - \hat{\mathbf{f}}, \mathbf{f}_1 - \hat{\mathbf{f}}, \dots, \mathbf{f}_n - \hat{\mathbf{f}}) \quad (35)$$

$$\hat{\mathbf{f}} = \frac{1}{n} \sum \mathbf{f}_i \quad (36)$$

for a set of  $n$  training images  $\mathbf{f}_i$  of the different objects from the data base. The eigenvectors  $\varphi_l$  that correspond to the  $k$  largest eigenvalues of  $\mathbf{Q}$  then form the matrix  $\Phi$  by

$$\Phi = (\varphi_1, \varphi_2, \dots, \varphi_k)^T \quad (37)$$

Instead of using one eigenspace for all object classes there exist also approaches that estimate for each object class  $\Omega_\kappa$  a transformation matrix  $\Phi_\kappa$  using only images  $\mathbf{f}_i$  from class  $\Omega_\kappa$ . In the following we will use one eigenspace for all classes.

After computing the transformation matrix  $\Phi$ , each training image  $f_i$  is projected into the eigenspace. The resulting feature vector  $c_i = \Phi f_i$  is stored together with the class label and sometimes pose parameters of the object in image  $f_i$ . During classification an image  $f$  is projected into the eigenspace and the decision is made for that class (and pose) for which the stored feature vector  $c_i$  has minimum distance to the vector  $\Phi f$ . Sometimes curves are fitted to the discrete positions of the feature vectors  $c_i$  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  $p(c|\Omega_\kappa)$  for each object class. Although there exists a very promising approach for probabilistic principal component analysis that results directly in the desired densities [Tipping and Bishop, 2000], for simplicity our implementation follows the approach of [Borotschnig *et al.*, 1998].

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

$$p(c|\Omega_\kappa) = p(\Phi f|\Omega_\kappa) = N(\mu_\kappa, \Sigma_\kappa^{-1}) \quad (38)$$

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

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

$$p(c|\Omega_\kappa, \mathbf{a}) = N_{\mathbf{a}}(\mu_\kappa, \Sigma_\kappa^{-1}) \quad . \quad (39)$$

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. (26). In our case  $m = 776$  and  $n = 9$ .

## 5 Related Work: Information Theory in Computer Vision

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 [Viola and Wells III, 1997], view point selection in object recognition [Schiele and Crowley, 1998] and feature extraction [Fisher and Principe, 1997].

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 [Schiele and Crowley, 1998]. The authors present an active object recognition scheme based on the transinformation 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 [Deinzer *et al.*, 2000] 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 [Borotschnig *et al.*, 1998] 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 [Viola and Wells III, 1997] 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 probability density functions 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 [Viola, 1995].

In [Fisher and Principe, 1997] 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 [Viola and Wells III, 1997] they represent the continuous probability density functions 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 [Niemann, 1990]. Related to our work the approach in [Fisher and Principe, 1997] 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 [Fox *et al.*, 1998], active view point selection for object recognition [Arbel and Ferrie, 1999], and sensor planning for active object search [Ye, 1997].

The most rigorous application of information theory in image processing and computer vision can be found in [Huck *et al.*, 1996]. 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 [Huck *et al.*, 1996] deals as a perfect 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 density functions and Monte Carlo sampling is new.





Figure 8: Test set: seven toy manikins seen from behind.

resolution of the images is  $256 \times 256$ .

The motivation for performing such synthetic zooming was the following. We wanted to examine the behavior of the proposed action selection mechanism in a controlled environment. Especially, we would predict that

- the recognition rate depends on the noise during zooming
- the quality of camera parameter selection depends on the number of training examples
- the number of trials to recognize the objects increases if the accuracy in camera parameter adjustment is low
- the system in general tends to choose lower resolution images if the accuracy in camera parameter adjustment is low.

Table 1 and Figure 9 support the claims above. In Figure 9 the recognition rate for the set of seven toy manikins is plotted against the simulated noise level (Gaussian noise with  $1 \leq \sigma^2 \leq 18$ , measured in pixels) in the pan-tilt movement of the camera for different numbers of training examples per object and pan-tilt-focal length position (50–10000 trials). One can see that for small noise levels ( $\sigma^2 \leq 2$ ) the recognition rate is 100% even for a small number of training examples, although only the mean gray value has been taken as feature.

The recognition rate is decreased as the noise level is increased and as the number of training examples is reduced. Both behaviors would have been expected, because large noise prevents the system from zooming close to the object, since then the probability of viewing the desired position at the object is decreased. This can also be observed in Table 1. Figure 9 also shows that with fewer training examples it is less likely that the current noise level can be estimated reliably. This is especially noticeable when only 50 images are used during training for each pan/tilt/zoom position.

In Table 1 the chosen zoom positions during the sequential decision process are shown depending on the noise level. As expected, with increasing noise level the system does not rely on its zooming ability, because viewing the desired part of the object at high resolution is more unlikely. Thus, the system uses more overview images, especially while being uncertain about the true object. Also one can observe that although the number of overview images ( $n = 2$ ) is increased, the number of detailed inspections ( $n = 16$ ) remains at a certain level. The reason is that if the system has ruled out most of the objects in most cases a final verification is done with a close-up view of the object. The last column in Table 1 shows the increased difficulty for the classification while the noise level is increased. For a noise level of  $\sigma^2 = 18$  in almost all cases the system needs the

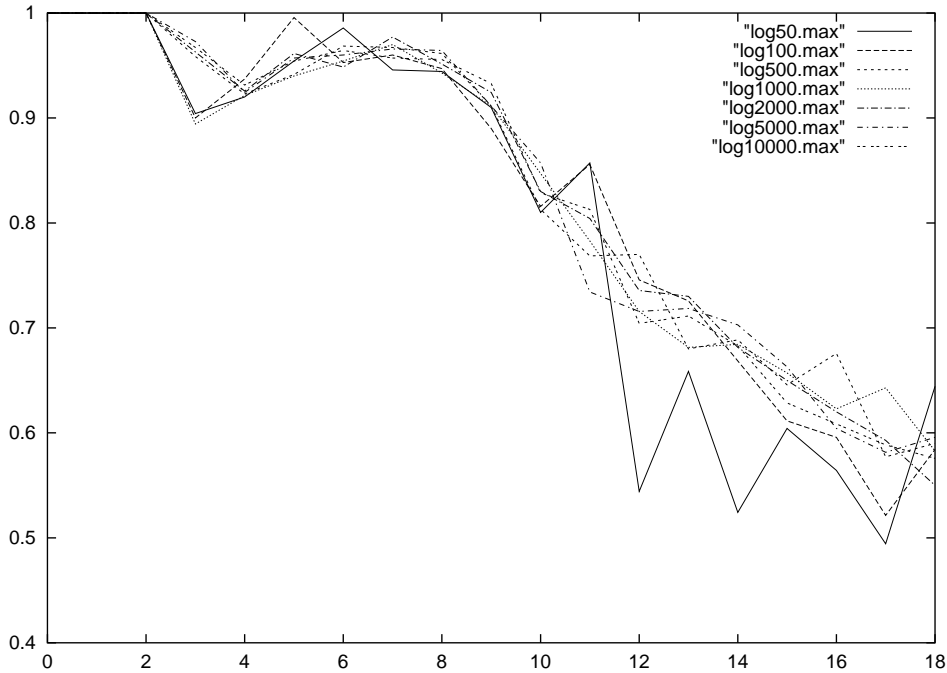


Figure 9: Recognition rate ( $y$ -axis) at different noise levels ( $x$ -axis, Gaussian noise with  $\sigma = 0, \dots, 18$ ) for the camera movement. The plots show the result for different number of training examples for each pan/tilt/focal length position, ranging from 50 to 10000.

maximum number of views (10) to come to a decision. In comparison to that for a noise level of  $\sigma^2 = 2$  on average 2.2 views are sufficient to classify the objects with 100% accuracy.

To emphasize the tremendous impact of noise on this problem imagine a noiseless environment, i.e. the changes in camera settings as well as the imaging are noiseless. Then the optimal sequence would consist of imaging three a pre-selected individual pixels to distinguish all objects from each other. The sequential decision process then (by training or knowledge) would look like a static decision tree.

## 6.2 Real Parameter Selection

In this section classification results from experiments with a real pan/tilt/zoom camera are presented. In Figure 10 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 10 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 10 (up to our level of discretization).

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:

noise $\sigma^2$	$n = 16$	$n = 8$	$n = 4$	$n = 2$	views
1	79.5	20.4	0	0	2.2
2	68.6	26.1	5.2	0	2.8
3	68.9	23.6	8.0	0	3.2
5	29.4	12.7	57.8	0	5.6
8	36.5	8.0	55.4	0	6.6
10	36.4	14.8	48.7	0	8.0
13	10.9	29.7	30.7	28.7	9.3
15	16.7	25.3	14.0	44.0	9.5
18	13.6	0	9.8	76.4	9.9

Table 1: Percentage of trials the zoom factor  $n$  is chosen depending on the (zero mean Gaussian) noise level. The value  $n = 16$  is the longest focal length,  $n = 2$  the shortest. With increasing noise the system tends to use more overview images ( $n = 2$ ) and also the mean number of views needed for classification is increased (last column). The number of detailed inspections decreases also with noise, but remains at a certain level, since detailed inspection is used in late classification stages, when most of the objects are already ruled out.

- 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 11. 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.

During training, the different densities in (9) 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 6):

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.

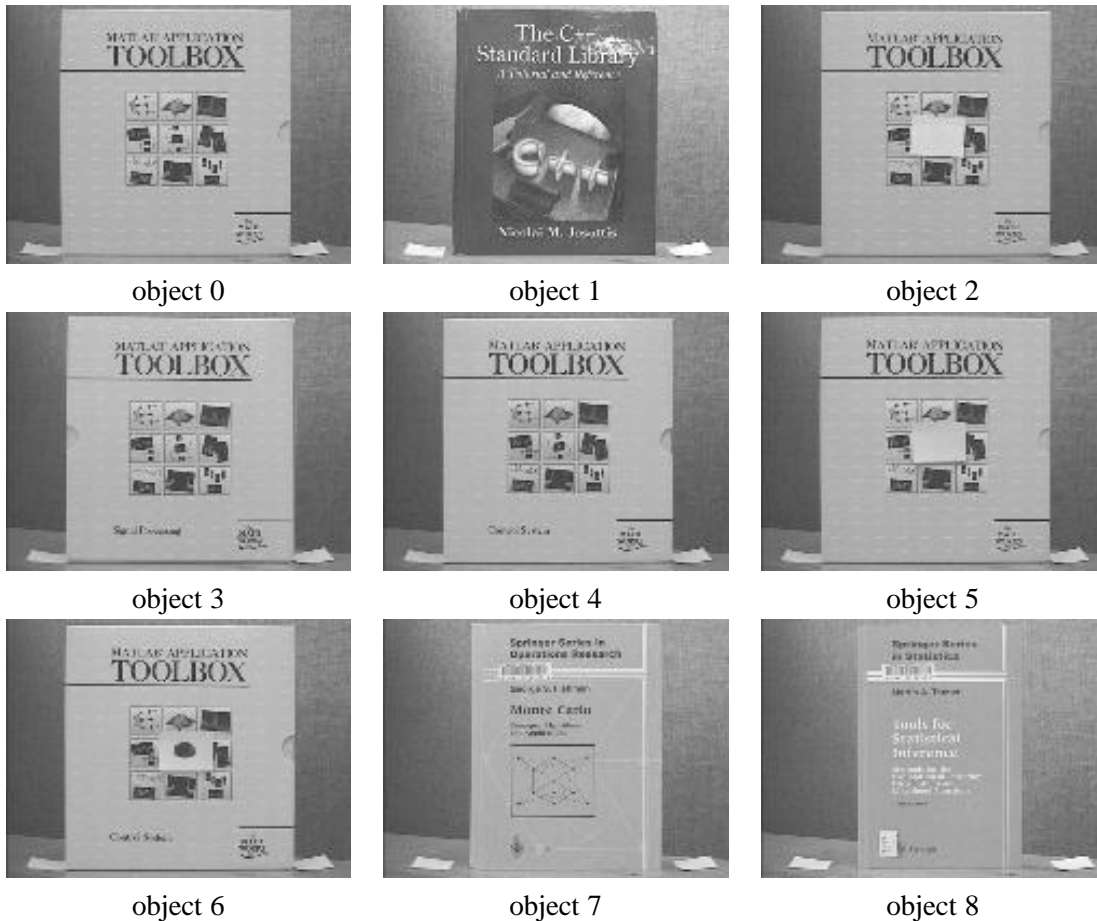


Figure 10: Data set for classification using zoom planning.

4. Bayes decision: Bayes formula (eq. (12)) 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 11 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.

In Figure 12 one example of successful active recognition for object number 3 is shown. The top row shows the sequence of observed images, the graphic under the images shows the change in the belief state, starting from a uniform distribution. After five views the belief for object 3 exceeds the probability 0.9 resulting in a decision for the right object. Here and in the following figures



Figure 11: Range in focal length: Left, shortest focal length. Right, longest focal length.

showing images taken during the sequential decision process the reader must be aware that the information processed by the automatic process consists of one of eight scalar integer numbers — the quantized mean gray value of the image. One can see that sequentially the other similar looking objects (object 0, 3 and 6) are ruled out by selecting the right gaze and focal length. Also worth noting is the fact that although no structural information is contained in the last captured image the extracted feature, i.e. the mean gray value, contains the necessary information for classification. At an earlier stage of the sequential decision process, i.e. for example, before looking to the center of the books, this might not be the case. In Figure 13 another experiment is depicted. Besides the change in belief state for the 9 classes one can also see the change in entropy of the distribution over 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 13 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 13, 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. 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.

A higher noise in the camera parameter control has been assumed in the experiment presented in Figure 14. One can see that in this case the entropy never increases at the cost that the decrease in uncertainty is dramatically reduced and also the final decision is not as unambiguous as for the experiment shown in Figure 13. Regardless, the maximum a posteriori decision after the last view returns the right class, i.e. object number 6. Another interesting observation in Figure 14 is that the system seems to use redundant information in views 4 and 5. But the two images are taken to compute the a posteriori probability using different a priori probabilities.

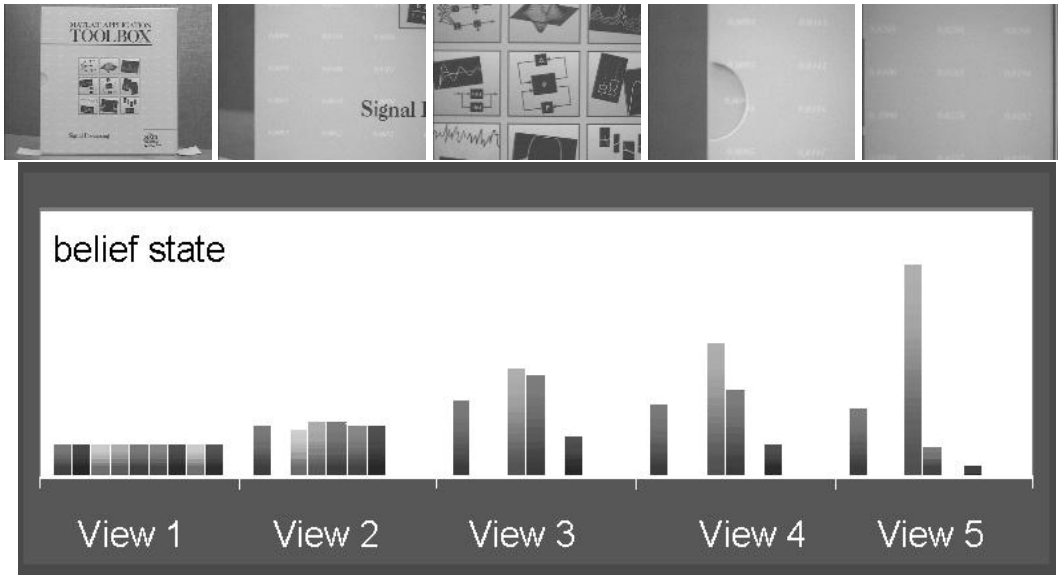


Figure 12: Object classification. Top row: sequence of views (object 3). Bottom row: change in belief state after interpreting each views.

Another experiment with object number 0 can be found in Figure 15 together with the change in belief and entropy. Again, the system returns the correct class after sequential gaze control.

In Table 2 the recognition results for the nine objects are shown. 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

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

mean gray value (compare also Figure 10), Thus, both objects have been taken 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 gaze control (third row in Table 2) 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

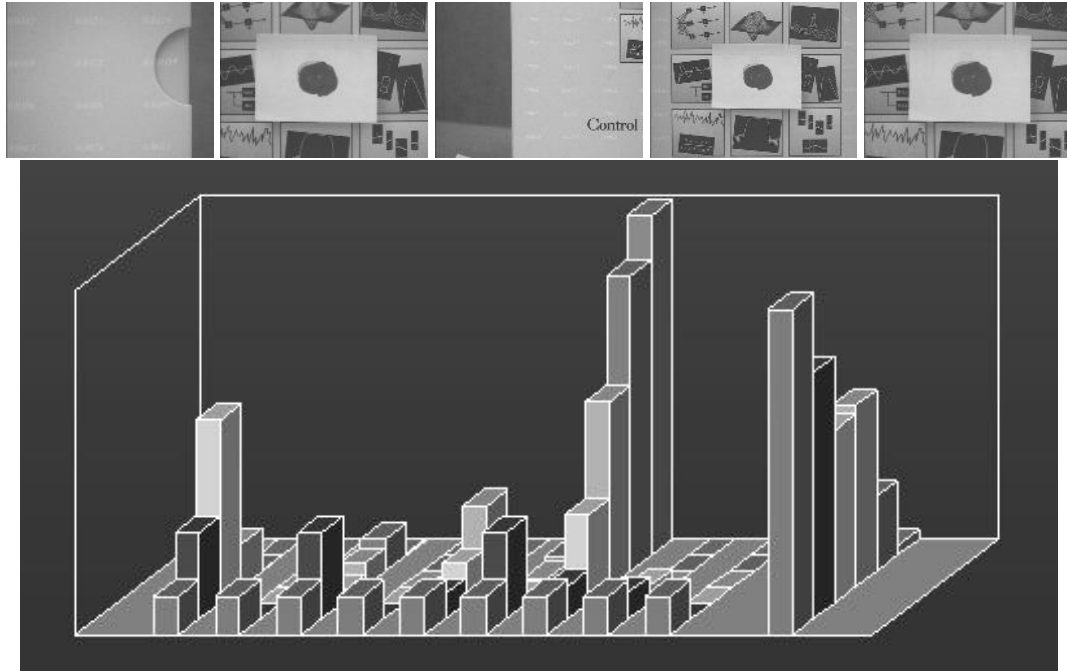


Figure 13: Object recognition using gaze control for object number 6 (the initial overview has been left out). Bottom: the change in belief state over time (from left to right, object number 0 to 8) and the change in entropy in the belief state (farthest right bars). The z-axis indicates the time step in the sequential decision process.

the mean number of views is increased from 1 to 2 views (compare Table 3). 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 fails completely.

### 6.3 Experiment with Statistical Eigenspace Classifier and Differential Mutual Information

As before we used the data set of the nine books shown in Figure 10. In the training step for each pan/tilt/zoom position  $\mathbf{a}$  we took views from each object class  $\Omega_{\kappa}$  to compute the transformation

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

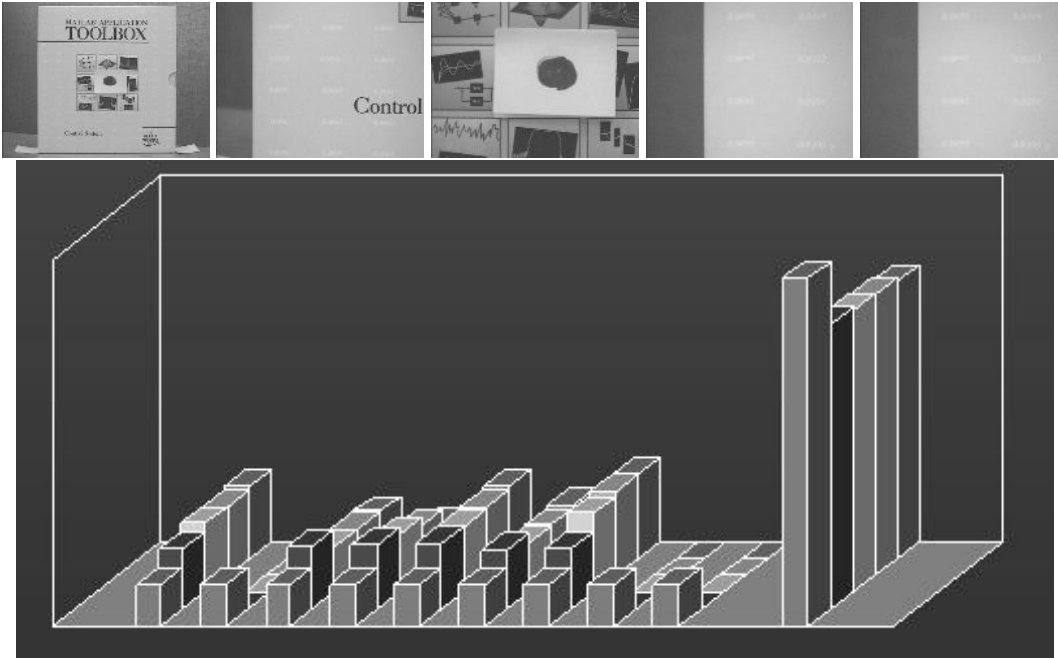


Figure 14: Object recognition with gaze control (object 6), where the noise level for pan-tilt-zoom positioning has been assumed to be high. Bottom: change in belief state and entropy. The reason the algorithm selects the identical views number 4 and 5 is the differing a priori probability in these views.

matrix  $\Phi_a$ . 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  $c_i$  have been 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.

In Table 4 the results for the view point planning strategy based on the maximum of mutual information is shown. The decision for the next view is made by Monte Carlo evaluation of the mutual information as described in the previous sections. For the Monte Carlo evaluation of the mutual information a total number of 1000 samples have been selected. One can see that 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 samples 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

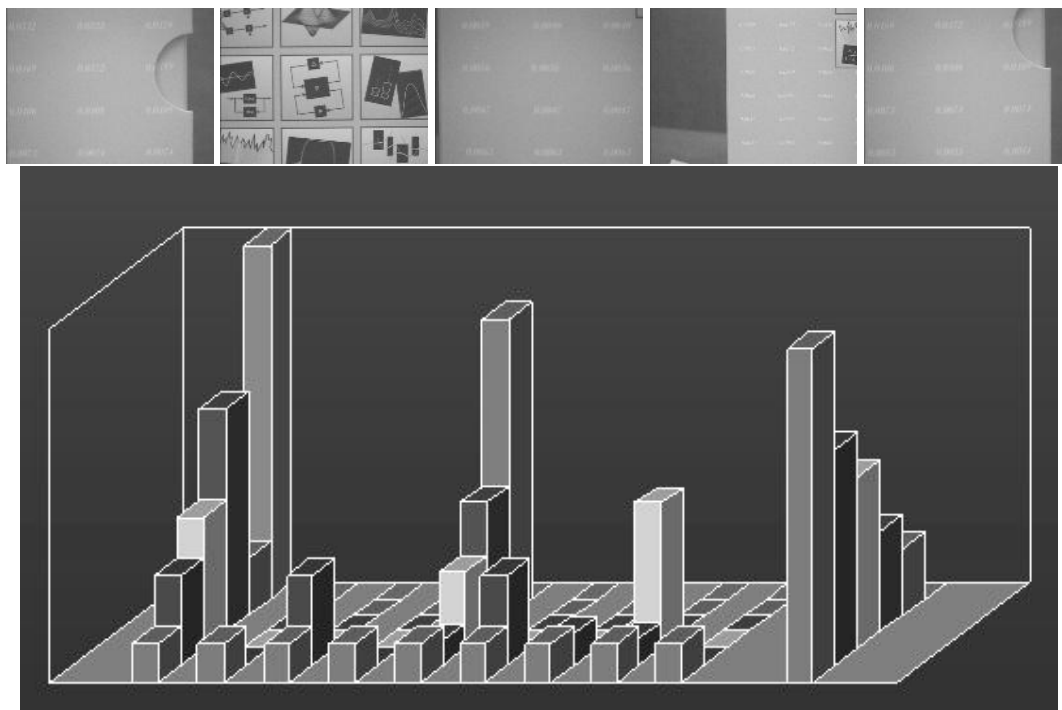


Figure 15: Object recognition with gaze control (object 0). Bottom: the change in belief state over time (from left to right, object number 0 to 8) and the change in entropy in the belief state (farthest right bars). The  $z$ -axis indicates the time step in the sequential decision process.

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 5 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

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		

Table 4: 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.

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 a priori 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

object	rec. rate	mean no. views	mean max. prob.
o0	83.4	10.9	0.61
o1	99.6	1.2	1.00
o2	62.4	10.8	0.65
o3	76.0	10.8	0.64
o4	66.6	11.0	0.56
o5	68.2	10.9	0.57
o6	76.7	10.7	0.63
o7	100.0	2.5	0.97
o8	100.0	2.4	0.97
average	81.4		

Table 5: 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.

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 have chosen 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.

For object recognition we have chosen a rudimentary state estimator based on the mean gray value in the captured image and using 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 strategy 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 probability density functions (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.

Nevertheless some points must be noted. First, our approach is completely embedded in a statistical framework. This means that assumptions for the underlying distributions must be made and verified and the estimation of the parameters of the densities is not a trivial problem, especially in higher dimensional spaces (state, feature, and action). The near future will show whether or not we can handle this curse of dimensionality when we map the ideas for continuous representation, estimation and evaluation of reward functions [Deinzer *et al.*, 2000] to the problem of representing and maximizing the mutual information. 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 [Cohn *et al.*, 1996].

Furthermore, in our future work we will apply a more general approach for representing probability density functions of random vectors, the so-called Parzen window density estimation. In [Viola, 1995; Viola and Wells III, 1997] 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.

## 8 Acknowledgments

The authors like to acknowledge Jochen Triesch and Robbie Jacobs for the discerning reviews of the manuscript, and Brian Madden for his detailed comments and discussions and for his pointer to the work of F. Huck.

Joachim Denzler personally would like to thank the people of the Computer Science Department of the University of Rochester for providing a pleasant and convenient environment for performing his research program, funded by the German Science Foundation (DFG) under grant DE 735/1.

This research was also partially supported by NSF grant EIA-9972881 and CAT/NYSSTF grant EEC-9813002.

## A Information Theory

In information theory five parts are important (compare also Figure 3 on page 5.) The main focus will be on the following three parts, since they are important for the problems tackled later on:

1. The *information source* produces a message, which has to be transmitted. These messages might be discrete or continuous.
2. The *channel* is the medium for transmitting the message produced by the information source, and which is only in the ideal case noiseless. Usually, a *noise source* disturbs the channel and changes the message.
3. The *destination* receives the message produced by the information source and sent over the channel.

The remaining two parts, the transmitter and receiver, are not important in the following, so that they are not further discussed.

Let us consider for the time being a discrete source. Discrete means that the message sent is chosen from a set of messages  $\mathcal{S} = \{s_1, \dots, s_n\}$ . Each message  $s_i \in \mathcal{S}$  is sent with probability  $p_i = p(s_i)$ . A commonly used example for an information source is natural written language, such as English or German. The set of messages consists of the letters of the alphabet; the probabilities are language dependent.

The channel, the transmission medium, is disturbed by the noise source, so that not in every case the received message  $r_i$  corresponds to the sent one. The noise in the transmission is mathematically described by a probability distribution  $p(r_i|s_i)$ . Finally, at the receiver side one is interested in the sent message, given the received one, i.e.  $p(s_i|r_i)$ .

In [Shannon, 1948] this whole process has been investigated from the information content point of view. The *entropy* of a stochastic process (in our case the source)

$$H = -K \sum_{i=1}^n p_i \log p_i$$

measures how much choice is involved in the selection of the event or how uncertain we are of the outcome of that event. The entropy has the following important general properties:

- $H$  is continuous in the  $p_i$ .
- If all events  $s_i$  are equal probable, i.e.  $p_i = \frac{1}{n}$ , then  $H$  is a monotonically increasing function of  $n$ .
- If a choice is broken down into two successive choices, the original  $H$  is the weighted sum of the individual values of  $H$ .

Further on, one can derive more specific properties:

- $H = 0$  if and only if all the  $p_i$  but one are zero, this one having value one.

- The maximum of  $H$  for a given  $n$  is equal to  $\log n$ , and is reached for  $p_i = \frac{1}{n}$ . This is the most uncertain situation.
- Any change toward equalization of the probabilities  $p_i$  increases  $H$ , more generally, if an averaging operation

$$p'_i = \sum_j a_{ij} p_j$$

is performed, with  $\sum_j a_{ij} = \sum_i a_{ij} = 1, a_{ij} > 0$   $H$  is increased, except the averaging operation corresponds to a permutation of the  $p_i$ , where then  $H$  remains the same.

- For the continuous case, similar properties can be derived. One important difference between discrete and continuous distributions is, that in the continuous case the entropy measures the randomness relative to the chosen coordinate system. This means, that if the coordinates are changed the entropy also changes in general. This property will be looked at later on in the case of tracking a moving object with  $n$  cameras.

For the source–channel–destination model some other quantities can be defined:

1. the entropy of the joint event  $s, r$  is defined as

$$H(s, r) = - \sum_{i,j} p(s_i, r_j) \log(s_i, r_j)$$

with  $H(s, r) \leq H(s) + H(r)$ .

2. the conditional entropy of  $r$  given  $s$ ,  $H(r|s)$ , is the average of the entropy of  $r$  for each value of  $s$  weighted according to the probability of getting that particular  $s$ , i.e.

$$H(r|s) = - \sum_i p(s_i) \sum_j p(r_j|s_i) \log(r_j|s_i) = - \sum_{i,j} p(r_j, s_i) \log(r_j|s_i) .$$

It can be shown that  $H(r|s) = H(r, s) - H(s)$  and following from that that  $H(r) \geq H(r|s)$ . An interpretation of this relationship is that on average, data helps you in guessing the outcome of  $r$ , because it does not increase uncertainty. The information contained in  $r$  given  $s$  remains the same if and only if the two random events  $r$  and  $s$  are independent.

3. Chain rule for entropy: From the above relations it is easy to derive the general chain rule, i.e.

$$H(r, s) = H(r) + H(s|r) = H(s) + H(r|s) \quad .$$

4. Using the channel model the conditional entropy  $H(s|r)$  is also called equivocation, the entropy  $H(r|s)$  is denoted as dissipation.
5. The mutual information between  $r$  and  $s$ , also called the transinformation of the channel is defined as

$$I(s; r) = H(s) - H(s|r) = H(r) - H(r|s) \geq 0 \quad .$$

It measures the average reduction in uncertainty about  $s$  that results from learning the value of  $r$  or vice versa. Another interpretation is that the mutual information measures the amount of information that  $s$  conveys about  $y$ .

## B Information Theory and Related Topics

The following summary of important facts, definitions, theorems and results from information theory are taken mainly from two books [Cover and Thomas, 1991; Jumarie, 1990], a preliminary book version [McKay, 2000], and the original work of Shannon [Shannon, 1948].

### B.1 Definitions

**Definition 1 (discrete entropy)** *The entropy  $H(X)$  of a discrete random variable  $X$  is defined by*

$$H(X) = - \sum p(x) \log p(x) \quad .$$

**Definition 2 (joint entropy)** *The joint entropy  $H(X,Y)$  of a pair of discrete random variables  $(X,Y)$  with a joint distribution  $p(x,y)$  is defined as*

$$H(X, Y) = - \sum \sum p(x, y) \log p(x, y) \quad .$$

**Definition 3 (conditional entropy)** *If  $(X, Y) \sim p(x, y)$  then the conditional entropy  $H(Y|X)$  is defined as*

$$\begin{aligned} H(Y|X) &= \sum p(x) H(Y|X = x) \\ &= - \sum p(x) \sum p(y|x) \log p(y|x) \\ &= - \sum \sum p(x, y) \log p(y|x) \quad . \end{aligned}$$

*In other words, the conditional entropy is the expected value of the entropies of the conditional distributions, averaged over the conditioning random variable*

**Definition 4 (Kullback Leibler distance)** *The relative entropy or Kullback Leibler distance between two probability mass function  $p(x)$  and  $q(x)$  is defined as*

$$D(p||q) = \sum p(x) \log \frac{p(x)}{q(x)} \quad .$$

**Definition 5 (mutual information)** *The mutual information  $I(x; y)$  is the relative entropy between the joint distribution  $p(x, y)$  and the product distribution  $p(x)p(y)$*

$$I(X; Y) = \sum \sum p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \quad .$$

**Definition 6 (conditional mutual information)** *The conditional mutual information of random variables  $X$  and  $Y$  given  $Z$  is defined by*

$$I(X; Y|Z) = H(X|Z) - H(X|Y, Z) \quad .$$

**Definition 7 (conditional relative entropy)** *The conditional relative entropy  $D(p(y|x)||q(y|x))$  is the average of the relative entropies between the conditional probability mass functions  $p(y|x)$  and  $q(y|x)$  averaged over the probability mass function  $p(x)$ , i.e.*

$$D(p(y|x)||q(y|x)) = \sum p(x) \sum p(y|x) \log \frac{p(y|x)}{q(y|x)} \quad .$$

## B.2 Theorems and Lemmas

**Lemma 1 (positivity)**  $H(X) \geq 0$  .

**Theorem 1 (chain rule)**  $H(X, Y) = H(X) + H(Y|X)$  .

**Corollary 1**  $H(X, Y|Z) = H(X|Z) + H(Y|X, Z)$

**Theorem 2 (mutual information and entropy I)**

$$\begin{aligned} I(X; Y) &= \sum \sum p(x, y) \log \frac{p(x, y)}{p(x)p(y)} \\ &= \sum \sum p(x, y) \log \frac{p(x|y)}{p(x)} \\ &= - \sum \sum p(x, y) \log p(x) + \sum \sum p(x, y) \log p(x|y) \\ &= - \sum p(x) \log p(x) - \left( - \sum \sum p(x, y) \log p(x|y) \right) \\ &= H(X) - H(X|Y) . \end{aligned}$$

**Theorem 3 (symmetry of mutual information)**

$$I(X; Y) = I(Y; X) = H(Y) - H(Y|X) = H(X) - H(X|Y) .$$

**Theorem 4 (mutual information and entropy II)**

$$I(X; Y) = H(X) + H(Y) - H(X, Y) .$$

**Theorem 5 (chain rule for entropy)** Let  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ . Then

$$H(X_1, X_2, \dots, X_n) = \sum H(X_i | X_{i-1}, \dots, X_1) .$$

**Theorem 6 (chain rule for information)**

$$I(X_1, X_2, \dots, X_n; Y) = \sum I(X_i; Y | X_{i-1}, \dots, X_1) .$$

**Theorem 7 (chain rule for relative entropy)**

$$D(p(x, y) || q(x, y)) = D(p(x) || q(x)) + D(p(y|x) || q(y|x)) .$$

**Theorem 8 (Jensen's inequality)** If  $f$  is a convex function and  $X$  is a random variable, then

$$E[f(x)] \geq f(E[X]) .$$

*Proof:* by induction and the definition of convexity.

**Theorem 9 (information inequality)**  $D(p||q) \geq 0$  and  $I(X; Y) \geq 0$ ; latter one follows from the first one. The same is true for the conditioned versions. *Proof:*  $\log$  is a convex function and Jensen's inequality.

**Theorem 10**  $H(X) \leq \log |A|$ , where  $A$  is the range of  $X$ . *Proof:* compute the relative entropy between  $u(x) = \frac{1}{|A|}$  and  $p(x)$  and make use of the positivity of the relative entropy.

**Theorem 11 (conditioning reduces entropy)**  $H(X|Y) \leq H(X)$  .

**Theorem 12 (Independence bound on entropy)** Let  $X_1, X_2, \dots, X_n$  be drawn according to  $p(x_1, x_2, \dots, x_n)$ . Then

$$H(X_1, X_2, \dots, X_n) \leq \sum H(X_i) \quad .$$

*Proof:* chain rule for entropy and conditioning reduces entropy.

**Theorem 13 (log sum inequality)** For non-negative numbers  $a_1, a_2, \dots, a_n$  and  $b_1, b_2, \dots, b_n$

$$\sum a_i \log \frac{a_i}{b_i} \geq \left( \sum a_i \right) \log \frac{\sum a_i}{\sum b_i}$$

and equality if  $\frac{a_i}{b_i} = \text{const}$ .

**Theorem 14**  $D(p||q)$  is convex in the pair  $(p, q)$ , i.e.

$$D(\lambda p_1 + (1 - \lambda)p_2 || \lambda q_1 + (1 - \lambda)q_2) \leq \lambda D(p_1 || q_1) + (1 - \lambda)D(p_2 || q_2)$$

for  $0 \leq \lambda \leq 1$ .

**Theorem 15 (concavity of entropy)**  $H(p)$  is a concave function of  $p$ , i.e.  $-H(p)$  is a convex function (see definition of convexity in the theorem above).

**Theorem 16** Let  $(X, Y) \sim p(x, y) = p(x)p(y|x)$ . The mutual information  $I(X; Y)$  is a concave function of  $p(x)$  for fixed  $p(y|x)$  and a convex function of  $p(y|x)$  for fixed  $p(x)$ .

**Theorem 17 (data processing inequality)** Let  $X \rightarrow Y \rightarrow Z$  form a Markov chain, i.e.  $p(x, y, z) = p(x)p(y|x)p(z|y)$ , then  $I(X; Y) \geq I(X; Z)$ , i.e. no processing of  $Y$  (deterministic or stochastic) can increase the information that  $Y$  contains about  $X$ . Equality is reached if  $X \rightarrow Z \rightarrow Y$  forms a Markov chain, and thus  $I(X; Y|Z) = 0$ .

**Corollary 2** If  $Z = g(Y)$  then  $I(X; Y) \geq I(X; g(Y))$ .

*Proof:*  $X \rightarrow Y \rightarrow g(Y)$  form a Markov chain.

**Corollary 3** If  $X \rightarrow Y \rightarrow Z$  form a Markov chain then  $I(X; Y|Z) \leq I(X; Y)$ .

### B.3 Informal Interpretations

**entropy** : the entropy of a random variable is a measure of the uncertainty of the random variable; it is a measure of the information required on the average to describe the random variable.

**mutual information** : is a measure of the amount of information than one random variable contains about another random variable. It is the reduction in the uncertainty of one random variable due to knowledge of the other,

**self information** since  $I(X; X) = H(X)$  the entropy is also called self information.

**conditioning reduces entropy** : knowing the outcome of one random experiment reduces the uncertainty in the outcome of another random experiment but only on the average, i.e. there might be certain outcomes that increase the uncertainty.

## C Convergence Proof of Sequential Decision Process

In this section we will prove that the sequential decision process will converge. Practically the process might be stopped as soon as only slight changes in the a posteriori probability are observed.

For the proof one theorem and two corollaries are necessary:

**Theorem 18** *The sequential decision process (Section 3.2) forms a Markov chain.*

**Proof.** Define a Markov chain  $(\mathcal{X}, \mathcal{O}, \mathcal{A}, p(x|a, x'))$ , with

- $\mathcal{X} = \{x_1, x_2, \dots, x_n\}$  being the states
- $\mathcal{O} = \{o_1, o_2, \dots, o_m\}$  being the observations
- $\mathcal{A} = \{a_1, a_2, \dots, a_l\}$  being the actions

and

$$p(x|a, x') = \sum_x p(x'|x, a, o)p(o|x, a) \quad (40)$$

being the state transition probability. The probability  $p(o|x, a)$  is called the observation probability, given state  $x$  and action  $a$ . The probability  $p(x'|x, a, o)$  is defined as

$$p(x'|x, a, o) = \begin{cases} 1 & \text{if } x' = \operatorname{argmax} p(x|a, o) \\ 0 & \text{otherwise} \end{cases} \quad (41)$$

where  $p(x|a, o)$  is the a posteriori probability for the states  $x$  after observing  $o$  under action  $a$ . In other words, the states  $\mathcal{X}$  represent the decision for a certain class  $x$  based on the a posteriori probability. A state transition is made if the maximum in the a posteriori probability switches from one class to the other. The action is defined in a similar way by

$$a = \operatorname{argmax}_a I(x; o|a) \quad (42)$$

using the maximum mutual information criterion. It can easily be verified that the above definitions result in a Markov chain.

**Corollary 4** Let  $p(x_n)$  and  $p'(x_n)$  be two distributions on a Markov chain at time step  $n$  then the Kullback–Leibler distance between these two distributions will never increase over time, i.e.  $D(p(x_n)||p'(x_n)) \geq D(p(x_{n+1})||p'(x_{n+1}))$ .

**Proof.** Let  $p(x_n, x_{n+1}) = p(x_n)r(x_{n+1}|x_n)$  and  $p'(x_n, x_{n+1}) = p'(x_n)r(x_{n+1}|x_n)$  be the corresponding joint mass function, where  $r(\cdot|\cdot)$  is the probability transition function for the Markov chain. Then by the chain rule for relative entropy, we have two expansions:

$$D(p(x_n, x_{n+1})||p'(x_n, x_{n+1})) \tag{43}$$

$$= D(p(x_n)||p'(x_n)) + D(p(x_{n+1}|x_n)||p'(x_{n+1}|x_n)) \tag{44}$$

$$= D(p(x_{n+1})||p'(x_{n+1})) + D(p(x_n|x_{n+1})||p'(x_n|x_{n+1})) \tag{45}$$

Since both  $p(x)$  and  $p'(x)$  are derived from the same Markov chain the conditional probability mass functions  $p(x_{n+1}|x_n)$  and  $p'(x_{n+1}|x_n)$  are equal to  $r(x_{n+1}|x_n)$  and hence

$$D(p(x_{n+1}|x_n)||p'(x_{n+1}|x_n)) = 0.$$

Now using the non–negativity of the relative entropy in general we have

$$D(p(x_n|x_{n+1})||p'(x_n|x_{n+1})) \geq 0$$

and thus

$$D(p(x_n)||p'(x_n)) \geq D(p(x_{n+1})||p'(x_{n+1})). \tag{46}$$

**Corollary 5** Relative entropy  $D(p(x_n)||p(x))$  between a distribution  $p(x_n)$  and a stationary distribution  $p(x)$  of the Markov chain decreases with  $n$ .

**Proof.** In the previous corollary  $p'(x_n)$  is any distribution over the Markov chain, i.e. the corollary holds also for a stationary distribution. By the definition of stationarity, i.e.  $p(x_n) = p(x_{n+1})$  we get

$$D(p(x_n)||p(x)) \geq D(p(x_{n+1})||p(x)). \tag{47}$$

This implies that any state distribution gets closer and closer to each stationary distribution as time passes. If the stationary distribution is unique than the lower bound is zero.

**Theorem 19 (Convergence of sequential decision process)** *The sequential decision process, consisting of maximum mutual information criterion for action selection and Bayes rule for computation of the a posteriori probability, converges.*

**Proof.** By Theorem 18 above the sequential decision process forms a Markov chain. Corollary 5 states that the distribution over the state of the Markov chain will converge toward to a point, where the distance to all stationary distributions of the Markov chain is minimized. If the stationary distribution is unique, the final distribution will be the stationary distribution itself.

In general it is difficult to prove whether there is only one stationary distribution in the Markov chain. If we have two copies of one object in the data set we would expect that we have two stationary distributions and the sequential decision process will converge toward the the point with minimum distance to both stationary distribution. Anyway, it remains open for now if there exist other stationary distributions than  $p(x) = (0, \dots, p_i = 1, \dots, 0)$  if object number  $i$  is in front of the camera.

## References

- [Arbel and Ferrie, 1999] 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.
- [Borotschnig *et al.*, 1998] 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.
- [Borotschnig *et al.*, 1999] 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.
- [Cohn *et al.*, 1996] D.A. Cohn, A. Ghahramani, and M.I. Jordan, “Active Learning with Statistical Models,” *Journal of Artificial Intelligence Research*, 4:129–145, 1996.
- [Cover and Thomas, 1991] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley Series in Telecommunications. John Wiley and Sons, New York, 1991.
- [Deinzer *et al.*, 2000] F. Deinzer, J. Denzler, and H. Niemann, “Viewpoint Selection - A Classifier Independent Learning Approach,” In *IEEE Southwest Symposium on Image Analysis and Interpretation*, 2000.
- [Fisher and Principe, 1997] J. Fisher and J.C. Principe, “A Nonparametric Method for Information Theoretic Feature Extraction,” In *DARPA Image Understanding Workshop*, New Orleans, 1997.
- [Fox *et al.*, 1998] D. Fox, W. Burgard, and S. Thrun, “Active Markov Localization for Mobile Robots,” Technical report, Carnegie Mellon University, 1998.
- [Huck *et al.*, 1996] 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.
- [Isard and Blake, 1996] 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.
- [Jumarie, 1990] G. Jumarie, *Relative Information*, Springer Series in Synergetics. Springer, Heidelberg, 1990.
- [Kaelbling *et al.*, 1998] 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.
- [Kalman, 1960] R.E. Kalman, “A New Approach to Linear Filtering and Prediction Problems,” *Journal of Basic Engineering*, pages 35–44, 1960.
- [Krebs *et al.*, 1998] B. Krebs, M. Burkhardt, and B. Korn, “Handling Uncertainty in 3D Object Recognition using Bayesian Networks,” In H. Burkhardt and B. Neumann, editors, *Computer Vision - ECCV 98*, pages 782–795, Berlin, Heidelberg, New York, London, 1998, Lecture Notes in Computer Science.

- [McKay, 2000] D.J.C. McKay, *Information Theory, Inference and Learning Algorithm*, unpublished, <http://www.cs.toronto.edu/~mackay/itprnn/book.html>, 2000.
- [Murase and Nayar, 1995] H. Murase and S. Nayar, “Visual Learning and Recognition of 3–D Objects from Appearance,” *International Journal of Computer Vision*, 14:5–24, 1995.
- [Niemann, 1990] H. Niemann, *Pattern Analysis and Understanding*, volume 4 of *Series in Information Sciences*, Springer, Berlin Heidelberg, 1990.
- [Russel and Norvig, 1994] S.J. Russel and P. Norvig, *Artificial Intelligence: A Modern Approach*, Prentice Hall, 1994.
- [Schiele and Crowley, 1998] B. Schiele and J.L. Crowley, “Transinformation for Active Object Recognition,” In *Proceedings of the Sixth International Conference on Computer Vision*, Bombay, India, 1998.
- [Shannon, 1948] C.E. Shannon, “A Mathematical Theory of Communication,” *The Bell System Technical Journal*, 27:397–423,623–659, 1948.
- [Sutton and Barto, 1998] R.S. Sutton and A.G. Barto, *Reinforcement Learning*, A Bradford Book, Cambridge, London, 1998.
- [Tanner, 1993] M.A. Tanner, *Tools for Statistical Inference*, Springer Verlag, London, Berlin, Heidelberg, New York, Paris, Tokyo, Hong Kong Budapest, 1993.
- [Tipping and Bishop, 2000] M.E. Tipping and C.M. Bishop, “Mixtures of Probabilistic Principal Component Analysers,” *Neural Computation*, page to appear, 2000.
- [Viola and Wells III, 1997] P. Viola and W.M. Wells III, “Alignment by Maximization of Mutual Information,” *International Journal of Computer Vision*, 24(2):137–154, 1997.
- [Viola, 1995] P.A. Viola, “Alignment by Maximization of Mutual Information,” Technical Report AI Technical Report No. 1548, MIT Artificial Intelligence Laboratory, 1995.
- [Ye, 1997] Y. Ye, “Sensor Planning for Object Search,” Technical Report PhD Thesis, Department of Computer Science, University of Toronto, 1997.