International Journal of Computer Vision 68(3), 267–287, 2006

DOI: 10.1007/s11263-005-4436-9

Efficient Discriminant Viewpoint Selection for Active Bayesian Recognition

CATHERINE LAPORTE AND TAL ARBEL
Centre for Intelligent Machines, McGill University, Montreal, Quebec, Canada
cathy@cim.mcgill.ca
arbel@cim.mcgill.ca

Received May 2, 2005; Revised August 18, 2005; Accepted August 19, 2005

First online version published in May, 2006

Abstract. This paper presents a novel viewpoint selection criterion for active object recognition and pose estimation whose key advantage resides in its low computational cost with respect to current popular approaches in the literature. The proposed observation selection criterion associates high utility with observations that predictably facilitate distinction between pairs of competing hypotheses by a Bayesian classifier. Rigorous experimentation of the proposed approach was conducted on two case studies, involving synthetic and real data, respectively. The results show the proposed algorithm to perform better than a random navigation strategy in terms of the amount of data required for recognition while being much faster than a strategy based on mutual information, without compromising accuracy.

Keywords: active vision, object recognition, pose estimation, Bayesian inference, efficient viewpoint selection

1. Introduction

The problem of object recognition is generally described as that of identifying an unknown object as one of a database of objects whose properties are known, based on observations of it made through a sensory device such as a camera. In this context, pose estimation refers to the additional problem of determining how the recognised object is positioned with respect to some global coordinate frame. The difficulties involved in solving these problems stem from the fact that more than one hypothesis can explain a particular observation. This is aggravated by the presence of uncertainty in the measurements and the influence of nuisance variables such as illumination. As such, it is not always possible to identify the object and estimate its pose with confidence from a single observation.

It was shown previously that these difficulties can be overcome by sequential recognition strategies that exploit measurements of the characteristics of the physical world as observed from multiple points of view (Seibert and Waxman, 1992; Chen and Chen, 2004; Gremban and Ikeuchi, 1994; Herbin, 1996; Kovačič et al., 1998; Arbel and Ferrie, 2001b). By incrementally performing inference on accumulated evidence, sequential recognition systems move the burden of coping with uncertainty and ambiguity away from the measurement and feature extraction steps which are inherently subject to varying amounts of noise. It has been shown that such strategies improve the performance of recognition systems (that is, with respect to single measurement strategies) by making them more robust with respect to ambiguities and less dependent on the particular feature extraction and matching algorithms used. Provided reasonable models of the world are available, over time, the combination of multiple observations leads to an unambiguous assessment of the data. This follows from the well-known information theoretic principle that knowledge about one random variable can only, on average, reduce the uncertainty in another random variable if the two variables are not independent (Cover and Thomas, 1991).

Acquiring data sequentially allows for an external control module to make decisions as to when to stop gathering data, or even *how* to acquire the next datum. In contexts where the acquisition of data is expensive, the number of measurements required to perform accurate recognition can be reduced by actively controlling the observation parameters (e.g., camera position, scale of a noise removal algorithm, etc.) such as to acquire measurements that are useful with respect to the task at hand. This idea is the essence of active vision systems (Aloimonos et al., 1988; Bajcsy, 1988).

In general, an active recognition system consists of three major components that interact in a cyclical manner throughout the recognition process. This structure is illustrated in Figure 1.

The *observation component* is responsible for acquiring sensory measurements of the world and, if needed, processing those measurements through some form of feature extraction. The observation process is tuned through a set of control parameters which can be adjusted by an external decision making agent. The *inference component* fuses the last observation with previously acquired data in order to make an assessment about the state of the world, taking into account the various sources of uncertainty. It summarises the current state of knowledge of the system. Based on the task requirements and on the current output of the inference component, the *observation selection component* controls the data acquisition process by choosing new parameter settings for the next observation to be made.

Despite its many benefits, the development of active recognition systems poses important challenges. First, it requires the integration of solutions to complex open problems. These problems include measurement parameterisation, evidence fusion and decision making. Second, an enormous amount of time and resources must be allocated to experimentation and data acquisition in order to validate the different components of active recognition strategies. Recent literature has focused on active vision problems where the goal was to minimise the amount of data needed to perform a recognition task; this goal was attained with some success with information theoretic criteria such as mutual information or expected loss of entropy (Borotschnig et al., 2000; Paletta et al., 2000; Denzler and Brown, 2002). A major drawback of several existing methods is their high computational complexity, which limits the dimensionality of the problems they can tackle.

This paper introduces a novel computationally efficient observation selection criterion that improves on traditional approaches. The proposed criterion associates high utility with observations whose outcome predictably facilitates distinction between pairs of competing hypotheses. The resulting algorithm is shown to have low complexity and lends itself to various simplifications. As such, it represents a low cost alternative to the popular solutions based on mutual information or average loss of entropy.

The proposed method is applied to a difficult, high dimensional problem, which poses a computational challenge to current active observation selection strategies. Furthermore, in contrast to many active object recognition approaches, this paper addresses the additional problem of disambiguating object pose in addition to object class. The dimensionality of the problem is further increased by the introduction of a nuisance parameter. The high dimensionality of this problem emphasises the strength of the proposed observation selection approach by showing how it provides a computationally efficient solution to a difficult set of problems.

The experimental results and analysis presented in this paper show that given accurate modeling information, the proposed approach reduces the number of observations required to achieve recognition with respect to a baseline random navigation approach, and that this reduction is similar to that obtained by maximisation of mutual information at a lower computational cost. This analysis also raises potentially problematic issues which appear when the models are of poor quality; namely, the active vision approach is more sensitive to modeling error than a random navigation approach. This mechanism is explored in depth, and a simple heuristic is proposed which largely alleviates the problem.

The remainder of the paper is structured as follows. Section 2 presents an overview of previous work in the field of active object recognition, emphasising the contribution made by this article. Section 3 then presents the sequential Bayesian evidence fusion framework which will be used to demonstrate the proposed active observation selection strategy, which is then described in Section 4. Finally, extensive experimental results based on synthetic and real data are discussed in Section 5, where an in depth analysis of the strengths and weaknesses of our approach is provided through comparisons with a baseline random navigation strategy and a popular observation selection strategy based on mutual information.

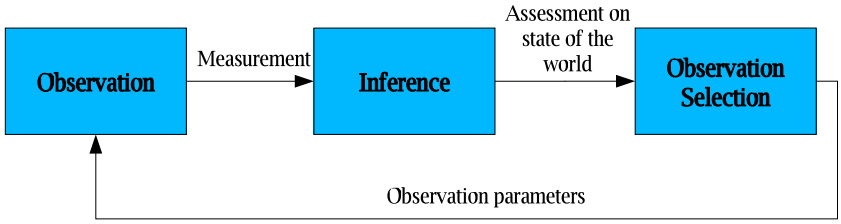


Figure 1. The structure of active recognition systems.

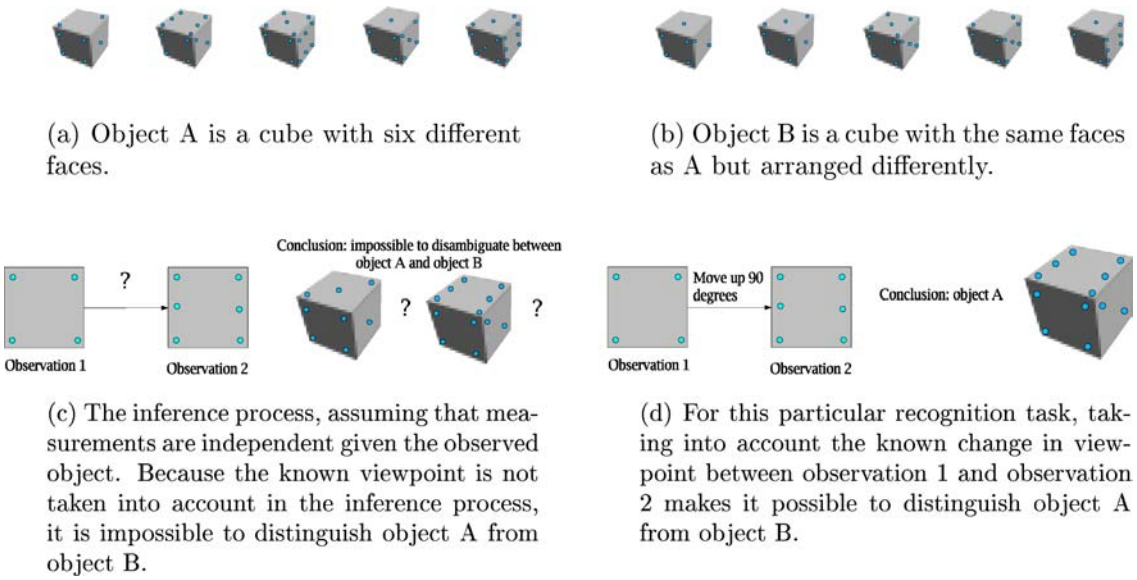
2. Previous Work

The active object recognition literature proposes various methods for effectively selecting observations. Several authors have argued in favour of off-line strategies that pre-compute a small set of characteristic views of the different objects of the database, chosen on the basis of their informativeness, as measured by some quantitative criterion. Examples of this include work by Dickinson et al. (1997), Schiele and Crowley (1998), Arbel and Ferrie (2001a, 2002) and Sipe and Casasent (2002). These strategies have the advantage of being conceptually and computationally simple. The on-line cost of decision making is negligible, and the cost of off-line training is usually quite low as well. However, these techniques typically rely on the assumption that prior estimates of one or two of the most probable hypotheses are sufficient to make a good decision as to the next observation, which is usually not the case in complex recognition problems. A strategy that attempts to distinguish the most probable hypothesis from all the others may be wasteful by not taking into account the fact that some of these have very low probability, thereby leading to the acquisition of redundant information. Another drawback of these methods is that they implicitly assume that the different hypotheses can be disambiguated on the basis of individual features. However, it may be the case that *structural* information brought by particular sequences of observations is needed (see for example Figure 2).

A more effective approach to the active recognition problem is to find the observation selection policy that minimises a measure of cost, such as the number of observations required to perform recognition with a particular level of confidence. The solution to this problem requires a global optimisation over all possible sequences of observations. The aspect graph based approach of Gremban and Ikeuchi (1994) attempts to provide such a solution by the construction of a heuristic "resolution tree" which is used to determine the shortest sensor displacement path leading to a unique solution to the recognition problem. An elegant approach for this kind of global optimisation is to cast the problem within the framework of partially observable Markov decision processes (Kaelbling et al., 1998), and solve it through dynamic programming or reinforcement learning (Sutton and Barto, 1998). Such approaches have been investigated by Darrell and Pentland (1995), Paletta and Pinz (2000), Ertin and Priddy (2002) and Geman and Jedynak (2001). Generally, global optimisation strategies are entirely computed off-line. The result is a decision policy in the form of a lookup table or a simple continuous function that is used to make decisions on-line. Thus, global optimisation strategies for viewpoint selection theoretically achieve optimality in a well defined sense and are computationally efficient on-line. Unfortunately, the amount of computations needed to estimate the optimal decision policy during training is usually prohibitive in any non-trivial task, even if these computations are performed off-line. Furthermore, as approximators must often be used in practice to represent continuous, high-dimensional belief spaces, dynamic programming and reinforcement learning algorithms may lead to solutions that are quite suboptimal (Sutton and Barto, 1998).

For these reasons, many authors prefer a local, myopic approximation of optimal behavior based on a one step lookahead measure of observation utility. The observation selection method proposed in this paper belongs to this category. The main difficulty in designing an observation selection strategy lies in the choice of a utility function. The literature proposes several alternatives, most of which involve maximising some measure of information gain. These approaches were shown to give good results in terms of the number of observations required for recognition. However, the evaluation of these criteria tends to become intractable as the dimensionality of the recognition problem grows.

Kovačič et al. (1998) present a strategy based on cluster analysis. During training, observations that yield similar measurements are clustered together



(a) Object A is a cube with six different faces.

(b) Object B is a cube with the same faces as A but arranged differently.

(c) The inference process, assuming that measurements are independent given the observed object. Because the known viewpoint is not taken into account in the inference process, it is impossible to distinguish object A from object B.

(d) For this particular recognition task, taking into account the known change in viewpoint between observation 1 and observation 2 makes it possible to distinguish object A from object B.

Figure 2. A recognition task where information about spatial structure is necessary to disambiguate between competing hypotheses.

using a method based on the minimal spanning tree. An active observation selection strategy is then devised whereby observation parameters are selected such that they maximally separate measurements which originally belonged to the same cluster. Conceptually, the work presented here is similar to their approach, with the fundamental difference being that the approach presented in this paper is based on a probabilistic model. In a certain sense, this allows for elements of the same cluster to be weighted differently in the calculations, according to the probability of the corresponding hypotheses.

Callari and Ferrie (2001) present a framework for active object recognition from range images. Shape primitives are fitted to the raw measurements, the parameters of which are then used to infer an object label. The combination of those two inference processes yields a probability over object labels given a set of raw measurements. Further views are chosen within a local neighbourhood of the current viewpoint by minimising the entropy of this distribution as it evolves over time. They show that their criterion effectively selects viewpoints that help distinguish the different objects by reducing the uncertainty in directions orthogonal to the decision boundaries. By making entropy calculations over a local region of viewpoint space, however, their strategy is prone to choosing suboptimal viewpoints which are local minima of the cost function.

Borotschnig et al. (2000) base their observation selection criterion on their probabilistic evidence fusion scheme for object recognition. Their strategy is to select the viewpoint that maximises the expected loss of entropy of the posterior distribution over object labels. Their strategy is computationally efficient because of the strong independence assumptions involved in their evidence fusion process. These assumptions make it impossible to distinguish objects which differ only in the spatial arrangement of their features (e.g., Figure 2).

Paletta et al. (2000) use the same viewpoint selection criterion as Borotschnig et al. (2000) with an inference model involving weaker independence assumptions, thereby solving the problems (Borotschnig et al., 2000) could not. This modification unfortunately renders on-line viewpoint selection a computationally expensive process. In order to render the observation selection strategy tractable, Paletta et al. approximate the utility function by constructing a lookup table off-line that maps a sequence of previously obtained measurements to the optimal viewpoint according to the chosen strategy. On-line, the obtained sequences of measurements are matched to their nearest neighbour in the lookup table and the corresponding observations are selected. This moves the computational burden to an off-line training phase. For large recognition problems, the time required for training becomes problematic.

Denzler and Brown (2002) emphasise the use of mutual information as a criterion for active observation selection. They develop a probabilistic formulation for evidence fusion and present mutual informa-

tion as the optimal measure of observation utility in the context of sequential decision making. Their paper presents experiments on both discrete and continuous measurement domains, and thereby illustrates how mutual information can be computed either exactly or using Monte-Carlo sampling. They claim that the distribution of measurements given viewpoint can be estimated off-line at low computational cost. However, this yields an *approximation* of the true distribution, which actually changes over time, as the probabilities of the possible states of the world fluctuate when more evidence is gathered.

Zhou et al. (2003) introduce a group of approaches to compute the immediate information gain that would result from measuring a particular feature in an active recognition context, given previously acquired information. They refer to their methodology as "conditional feature sensitivity analysis" because of its ability to take into account the dependencies between various types of measurements.

Previous research has focused on developing observation selection criteria that maximally reduce the number of observations needed for recognition. Particularly interesting and popular in this respect are information theoretic criteria such as mutual information (Denzler and Brown, 2002) or average loss of entropy (Borotschnig et al., 2000; Paletta et al., 2000). These have strong theoretical foundations, but are computationally expensive to evaluate.

In previous versions of this work (Laporte et al., 2004), a method inspired from linear discriminant analysis was suggested that was shown to overcome this problem by reducing the cost of decision making. This paper proposes a similar, but more general and rigorous strategy based on information theory, which has the same appealing computational properties. The method is applied to a more general recognition problem, where the light source is introduced as a nuisance variable.

3. Sequential Bayesian Recognition

The foundation of the active viewpoint selection approach presented in this paper lies in a Bayesian formulation of the recognition problem. The solution to the recognition problem is represented as a probability distribution over the possible hypotheses. This Bayesian framework, closely inspired from previous work (Arbel and Ferrie, 2001b; Borotschnig et al., 2000; Denzler and Brown, 2002; Paletta et al., 2000), allows for new data to be integrated with results inferred from prior data in an efficient way. Furthermore, the framework provides a natural framework for the development of information theoretic measures of observation utility. Note that the sequential Bayesian recognition strategy described in this section is used in all experiments described in Section 5, including those using random viewpoint selection, and that the costs attached to it are not particular to the choice of a particular observation selection method.

Consider a database of objects o_i, $i \in \{1, \ldots N_o\}$ whose characteristics are known prior to experimentation and a mobile camera facing an unknown object from this set whose class and pose are to be determined. The objects may be positioned in any of N_θ discrete poses defined according to a global reference frame. The observed scene may be illuminated by one of N_l different light sources, l. Let the camera measurement be parameterised by a feature vector $\mathbf{d}$, which depends on the identity o of the object, its pose θ, the light source l used to illuminate the scene and the viewing position $\mathbf{v}$. Under uncertainty, this relationship can be represented through a conditional probability density function $p(\mathbf{d} \mid o,\theta,l,\mathbf{v})$ whose parameters are assumed to be obtained from a physical model or estimated off-line from training data.[1] In the context of observation selection (Section 4), this distribution obtained from modeling is used as time context-independent information describing the appearance of any particular object.

Before describing the sequential Bayesian recognition strategy in more detail, an important note should be made about the chosen representation of objects. In the development of this work, a key requirement was to keep the evidence fusion and observation selection approaches decoupled from object representation as much as possible, as the choice of such a representation is typically a function of the nature of the objects in the database and available measurement modalities. Therefore a generic object representation was sought, which led to the form of probability density functions relating appearance to the problem variables (object class, pose and illumination) and control parameters (viewpoint). This merely requires that arbitrary features may be extracted in a repeatable manner from images of the object and as such, accounts for a broad range of object representations including appearance-based representations, interest-point methods, parametric shape primitives and aspect graphs.

The task of the inference engine is to recover the class o and pose θ of the observed object from a set of

measurements. The light source l used to illuminate the scene, in this case, acts as a nuisance variable whose estimation is not essential to the task, but still affects the observations in a tangible manner. Note that in a typical computer vision setting, illumination is a continuous variable. In this paper, the choice of illumination as a discrete nuisance variable is simply an example which facilitated experimentation. The formulation presented here can just as well be applied to other computer vision problems where discrete nuisance variables occur naturally.

Given a measurement $\mathbf{d}$, a known viewing position $\mathbf{v}$ and a prior distribution $P(o, \theta, l)$ over object class, object pose and light source, the probability of each class-pose-light source tuple is computed using Bayes' rule:

$$P(o, \theta, l \mid \mathbf{d}, \mathbf{v}) \propto p(\mathbf{d} \mid o, \theta, l, \mathbf{v})\, P(o, \theta, l), \quad (1)$$

safely assuming that the viewpoint is independent of the object, pose and light source under observation, i.e., $P(o, \theta, l \mid \mathbf{v}) = P(o, \theta, l)$. This independence assumption merely implies that given no data, the nature of the scene usually does not constrain possible starting viewpoints.

Different measurements may then be obtained by varying the observation parameters (i.e. the viewing position $\mathbf{v}$). It is assumed that subsequent measurements are independent of each other given the state of the world (defined by the object class, its pose and the light source) and the observation parameters (i.e. the viewing positions) that were used to make them; that is,

$$p(\mathbf{d}_t \mid o, \theta, l, \mathbf{d}_{t-1}, \mathbf{v}_t, \mathbf{v}_{t-1}) = p(\mathbf{d}_t \mid o, \theta, l, \mathbf{v}_t), \quad (2)$$

where $\mathbf{d}_t$ and $\mathbf{v}_t$ are the measurement and viewpoint at step t. This assumption leads to the following recursive Bayesian update rule:

$$P(o, \theta, l \mid \mathbf{d}_t, \mathbf{v}_t, \ldots, \mathbf{d}_1, \mathbf{v}_1) \propto \mathrm{p}(\mathbf{d}_t \mid o, \theta, l, \mathbf{v}_t)\, P(o, \theta, l \mid \mathbf{d}_{t-1}, \mathbf{v}_{t-1}, \ldots, \mathbf{d}_1, \mathbf{v}_1). \quad (3)$$

The posterior distribution $P(o, \theta, l \mid \mathbf{d}_t, \mathbf{v}_t, \ldots, \mathbf{d}_1, \mathbf{v}_1)$ provides contextual information which can be used later for inference (see remainder of this section) and/or observation selection (see Section 4). As the sensor is moved to new locations and more observations of an object are taken, a sequential recognition engine based on this evidence fusion scheme exploits the information provided by the appearance of the object and, more importantly, by its spatial structure (Paletta et al., 2000). Figure 2 illustrates a recognition task for which this structural information is necessary. This is in contrast with fusion strategies that assume independence between measurements given only the observed object (e.g. Arbel and Ferrie, 2001b; Borotschnig et al., 2000), which do not measure the temporal consistency between subsequent observations with respect to the database models (Paletta et al., 2000).

Finally, although the evidence fusion process always takes into account all variables, a probability distribution over the variables of interest (i.e. object class and pose only) can be obtained at any point in time by marginalising Eq. (3) over the light source nuisance variable:

$$P(o, \theta \mid \mathbf{d}_t, \mathbf{v}_t, \ldots, \mathbf{d}_1, \mathbf{v}_1) = \sum_{i=1}^{N_l} P(o, \theta, l_i \mid \mathbf{d}_t, \mathbf{v}_t, \ldots, \mathbf{d}_1, \mathbf{v}_1).$$

It is straightforward to generalise this process to a continuous illumination variable by replacing the sum with an integral. The strength of an approach such as this one is that based on this result, decisions can be made as to the nature of the observed scene (based, for instance, on the MAP object class and pose combination), or as to whether more data should be acquired, and if so, where to seek it, as will be shown in the next section.

4. Active Observation Selection

The object recognition and pose estimation problem is difficult, mainly because for certain choices of viewing positions $\mathbf{v}$, the observed data may be well explained by more than one hypothesis. As a result, the measurement distributions corresponding to the competing hypotheses are said to be similar for these particular choices of $\mathbf{v}$. A natural way to alleviate this difficulty is to present the inference engine with measurements that are inherently unambiguous with respect to the evidence already acquired. This can be achieved by choosing observation parameters such that, regardless of the true object and pose under observation, the most likely distributions of the resulting measurements are predictably dissimilar. That is, a shift in viewpoint must be chosen such that competing hypotheses will appear as different as possible such as to facilitate distinction. Considering the recognition task as a series of pair-

wise discrimination subtasks, and given a measure of dissimilarity $\Delta(p \parallel q)$ between two probability density functions p and q, the following general form is proposed as a criterion for the selection of a viewpoint $\mathbf{v}$ at step $t+1$:

$$\mathbf{v}^*_{t+1} = \underset{\mathbf{v}_{t+1}}{\operatorname{argmax}} \sum_{i=1}^{N_o} \sum_{j=1}^{N_\theta} \sum_{k=i}^{N_o} \sum_{m=m_{ijk}}^{N_\theta} P(o_i, \theta_j | \mathfrak{D}_t) P(o_k, \theta_m | \mathfrak{D}_t) \Delta(p(\mathbf{d}_{t+1} | o_i, \theta_j, \mathbf{v}_{t+1}, \mathfrak{D}_t) \parallel p(\mathbf{d}_{t+1} | o_k, \theta_m, v_{t+1}, \mathfrak{D}_t)), \tag{4}$$

where $\mathfrak{D}_t \equiv \{\mathbf{d}_t, \mathbf{v}_t, \ldots, \mathbf{d}_1, \mathbf{v}_1\}$ and

$$m_{ijk} = \begin{cases} j+1 & k = i \\ 1 & k > i. \end{cases}$$

An intuitive interpretation of this criterion is as follows: each term of the sum is achieving a pairwise comparison of two possible hypotheses (o_i, θ_j) and (o_k, θ_m). The hypotheses which are most likely account for most of the ambiguity in the state of knowledge of the system. Therefore, more effort is made to disambiguate likely hypotheses than unlikely hypotheses. The dissimilarity between the distribution of measurements as they would be obtained from viewpoint $\mathbf{v}_{t+1}$ is measured by the operator Δ, and this dissimilarity measure is weighted by the posterior probabilities associated with the compared hypotheses, conditioned on the evidence accumulated so far, $P(o_i, \theta_j \mid \mathfrak{D}_t)$ and $P(o_k, \theta_m \mid \mathfrak{D}_t)$. The pairwise dissimilarity operator measures the extent to which an observation is useful in disambiguating two hypotheses, and the weights measure the extent to which these hypotheses need to be disambiguated, based on how probable they are.

4.1. *The Jeffrey Divergence as a Measure of Dissimilarity*

The general form Eq. (4) does not make any assumptions about the nature of the dissimilarity measure Δ between two probability density functions. The choice of such a measure is therefore largely a design issue. The approach chosen in this paper is based on notions from information theory. An well-known information-theoretic measure of dissimilarity between two multivariate probability density functions p and q is the Kullback-Leibler (K-L) divergence, also known as the relative entropy (Cover and Thomas, 1991). For two multivariate probability densities p and q, the K-L divergence is defined as

$$D(p \,||\, q) = \int_{-\infty}^{\infty} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x}. \tag{5}$$

In information theoretic terms, $D(p \,||\, q)$ represents the amount of information lost if samples from the distribution p are assumed to be samples from q. It measures how poorly q models the distribution of data samples originating from a distribution p. This quantity is not symmetric ($D(p\|q) \neq D(q\|p)$) and as such, it is rather inconvenient to use as a dissimilarity measure. A simple alternative is to use instead the Jeffrey divergence, which is a symmetric measure of dissimilarity between two distributions defined as

$$\begin{aligned} J(p \parallel q) &= D(p \parallel q) + D(q \parallel p) \\ &= \int_{-\infty}^{\infty} p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})} d\mathbf{x} \\ &\quad + \int_{-\infty}^{\infty} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x})} d\mathbf{x}. \end{aligned} \tag{6}$$

The Jeffrey divergence is a measure of the expected loss of information incurred by type I and type II errors in processing data sampled from two different distributions. This can be viewed as a measure of how difficult it is to generate samples from those distributions such that they will be misinterpreted.

A more intuitive explanation of what the Jeffrey divergence measures is obtained by analysing the special case of the comparison between two Gaussian distributions p_1, p_2 with mean vectors and covariance matrices μ_1, μ_2, $\mathbf{C}_1$, $\mathbf{C}_2$. In this case, the Jeffrey divergence is given by

$$\begin{aligned} J(p_1 \parallel q_2) = \frac{1}{2} \big(&(\mu_1 - \mu_2)^T \big(\mathbf{C}_1^{-1} + \mathbf{C}_2^{-1}\big)(\mu_1 - \mu_2) \\ &+ \operatorname{tr}\big(\mathbf{C}_1^{-1}\mathbf{C}_2 + \mathbf{C}_2^{-1}\mathbf{C}_1 - 2\mathbf{I}\big)\big), \end{aligned} \tag{7}$$

where $\operatorname{tr}(\cdot)$ denotes the trace operator and $\mathbf{I}$ is the identity matrix of dimensions equal to those of $\mathbf{C}_1$ and $\mathbf{C}_2$. A detailed derivation of this result is provided as an appendix in (Laporte, 2004). When Eq. (7) is used within an observation selection criterion of the form given in Eq. (4), each term in the summation is identical to the measure of information gain proposed by MacKay (1992b) for selecting data points to perform optimal distinction between two models. The first term of Eq. (7) measures the distance between the means of the two distributions according to a scale defined by $\mathbf{C}_1$

and $\mathbf{C}_2$. The larger this distance, the more difficult it is to mistake a sample from one distribution for a sample from the second distribution. The second term is a function of the difference between the variances of the two distributions and accounts for what MacKay refers to as the "Occam factor" which penalises models with high variance in the context of Bayesian classification (MacKay, 1992a). In other words, models with significantly different variances are unlikely to be confused with one another by a Bayesian classifier.

Note that if the two covariance matrices are equal, then the Jeffrey divergence is equal to the Mahalanobis distance between the two densities. In previous versions of this work (Laporte et al., 2004), a closely related quantity inspired from linear discriminant analysis was suggested as a dissimilarity measure between probability density functions:

$$M(p_1, p_2) = (\mu_1 - \mu_2)^T (\mathbf{C}_1 + \mathbf{C}_2)^{-1} (\mu_1 - \mu_2). \quad (8)$$

In that work, it was assumed that the two distributions being compared were well represented by their means and covariance matrices, and the proposed dissimilarity measure was mainly an evaluation of the separability of the means of the appearance distributions. The use of the Jeffrey divergence as a dissimilarity measure is a more general approach in that it allows for comparison of arbitrary distributions, while taking into account all their parameters.

4.2. *Incorporating Prior and Contextual Knowledge*

The observation selection criterion described by Eq. (4) only attempts comparisons between different values of the variables of interest, i.e. the object class and pose pairs, not including nuisance variables (such as illumination) in their description. However, the raw information that is available to the observation selection component of the active recognition system is given in terms of relationships involving both variables of interest (o and θ) and the nuisance variable l. Some work must therefore be done to express the observation selection criterion Eq. (4) in terms of the available probability distributions $p(\mathbf{d} \mid o, \theta, \ l, \mathbf{v})$ (the prior information obtained from modeling) and $P(o, \theta, \ l \mid \mathfrak{D}_t)$ (the contextual information obtained by performing inference from visual data). Using the Jeffrey divergence as a measure of dissimilarity, Eq. (4) becomes,

$$\mathbf{v}^*_{t+1} = \underset{\mathbf{v}_{t+1}}{\operatorname{argmax}} \sum_{i=1}^{N_o} \sum_{j=1}^{N_\theta} \sum_{k=i}^{N_o} \sum_{m=m_{ijk}}^{N_\theta} P(o_i, \theta_j \mid \mathfrak{D}_t) P(o_k, \theta_m \mid \mathfrak{D}_t)$$

$$J(p(\mathbf{d}_{t+1}|o_i, \theta_j, \mathbf{v}_{t+1}, \mathfrak{D}_t) \parallel p(\mathbf{d}_{t+1}|o_k, \theta_m, \mathbf{v}_{t+1}, \mathfrak{D}_t)), \quad (9)$$

where $P(o, \theta|\mathfrak{D}_t)$ is obtained from Eq. (4) and $p(\mathbf{d}_{t+1}|o, \theta, \mathbf{v}_{t+1}, \mathfrak{D}_t)$ is obtained using the standard rules of probability and the independence assumption Eq. (2):

$$p(\mathbf{d}_{t+1} \mid o, \theta, \mathbf{v}_{t+1}, \mathfrak{D}_t) = \sum_{i=1}^{N_l} p(\mathbf{d}_{t+1} \mid o, \theta, l_i, \mathbf{v}_{t+1}) \frac{P(o, \theta, l_i \mid \mathfrak{D}_t)}{P(o, \theta \mid \mathfrak{D}_t)}. \quad (10)$$

The light source variable, which acts as a nuisance variable in the object recognition and pose estimation problem, was implicitly taken into account and marginalised away.

4.3. *Computational Considerations*

The evaluation of the Jeffrey divergences involved in Eq. (9) is a potential computational bottleneck. If the form of the statistical appearance model is defined by continuous probability density functions, computing the required integrals (see Eq. (6)) may be difficult in the presence of modeled nuisance variables, such as the light source l. In this case, Eq. (9) involves the computation of the Jeffrey divergence between mixture densities of the form (Eq. (10)), whose weights are determined by the probabilities associated with the possible values of l for each object class and pose pair. The K-L divergence (and, by extension, the Jeffrey divergence) between two mixture densities can generally not be expressed in closed form (Vasconcelos, 2001; Goldberger et al., 2003). The required integrals (see Eq. (6)) can be computed using methods such as numerical quadrature or Monte-Carlo sampling, but this is computationally expensive. Goldberger et al. (2003) used a quick and reasonably accurate upper-bound approximation to the K-L divergence which they used sucessfully in the context of content-based image retrieval. This work adopts this approximation to obtain a reasonably tight upper bound on the desired Jeffey

divergence:

$$
\begin{aligned}
&J(p(\mathbf{d}_{t+1}|o_i,\theta_j,\mathbf{v}_{t+1},\mathfrak{D}_t) \parallel p(\mathbf{d}_{t+1}|o_k,\theta_m,\mathbf{v}_{t+1},\mathfrak{D}_t))\\
&\leq \sum_{n=1}^{N_l} P(l_n|o_i,\theta_j,\mathfrak{D}_t)\log\frac{P(l_n|o_i,\theta_j,\mathfrak{D}_t)}{P(l_n|o_k,\theta_m,\mathfrak{D}_t)}\\
&+\sum_{q=1}^{N_l} P(l_q|o_i,\theta_j,\mathfrak{D}_t)D(p(\mathbf{d}_{t+1}|o_i,\theta_j,l_q,\mathbf{v}_{t+1})\\
&\parallel p(\mathbf{d}_{t+1}|o_k,\theta_m,l_q,\mathbf{v}_{t+1}))\\
&+\sum_{r=1}^{N_l} P(l_r|o_k,\theta_m,\mathfrak{D}_t)\log\frac{P(l_r|o_k,\theta_m,\mathfrak{D}_t)}{P(l_r|o_i,\theta_j,\mathfrak{D}_t)}\\
&+\sum_{s=1}^{N_l} P(l_s|o_k,\theta_m,\mathfrak{D}_t)D(p(\mathbf{d}_{t+1}|o_k,\theta_m,l_s,\mathbf{v}_{t+1})\\
&\parallel p(\mathbf{d}_{t+1}|o_i,\theta_j,l_s,\mathbf{v}_{t+1})). \qquad (11)
\end{aligned}
$$

Observe that the K-L divergences in the second term and the fourth term can be computed off-line as they only involve the statistical appearance model $p(\mathbf{d} \mid o,\theta,l,\mathbf{v})$. The remaining on-line computations are far less expensive than a perhaps more accurate Monte-Carlo sampling approach (Goldberger et al., 2003). With a non-varying light source, the approximation (Eq. (11)) is exact and the Jeffrey divergences in Eq. (9) can be entirely computed off-line from the available statistical appearance model.

For the purpose of comparison, consider two popular active observation selection criteria based on information theory chosen because they have been shown to yield good results in terms of reducing the number of observations needed for recognition. The first is based on the expected loss of entropy of the joint posterior distribution over object class and pose (Paletta et al., 2000):

$$
\begin{aligned}
\mathbf{v}^*_{t+1} = \underset{\mathbf{v}_{t+1}}{\operatorname{argmax}}\, &E\{H(o,\theta \mid \mathfrak{D}_t)\\
&-H(o,\theta \mid \mathbf{d}_{t+1},\mathbf{v}_{t+1},\mathfrak{D}_t)\}, \qquad (12)
\end{aligned}
$$

where $H(\cdot|\cdot)$ denotes the conditional entropy of a random variable (Cover et al., 1991). Typically, Eq. (12) is computed using Monte-Carlo methods to predict the outcome of the next evidence fusion step and requires $O(N_S N_o^2 N_\theta^2 N_l^2)$ operations to evaluate one viewpoint, where N_S is the number Monte-Carlo samples.

The second criterion is based on the mutual information between the new measurement $\mathbf{d}_{t+1}$ and the joint object class and pose, (o,θ) (Denzler and Brown, 2002):

$$
\begin{aligned}
\mathbf{v}^*_{t+1} = \underset{\mathbf{v}_{t+1}}{\operatorname{argmax}} &\sum_{i=1}^{N_o}\sum_{j=1}^{N_\theta} P(o_i,\theta_j|\mathfrak{D}_t)\\
&\times \int_{\mathbf{d}} p(\mathbf{d}_{t+1}|o_i,\theta_j,\mathbf{v}_{t+1},\mathfrak{D}_t)\\
&\times \log\frac{p(\mathbf{d}_{t+1}|o_i,\theta_j,\mathbf{v}_{t+1},\mathfrak{D}_t)}{\sum_{k=1}^{N_o}\sum_{m=1}^{N_\theta} p(\mathbf{d}_{t+1}|o_k,\theta_m,\mathbf{v}_{t+1},\mathfrak{D}_t)} d\mathbf{d}_{t+1}. \qquad (13)
\end{aligned}
$$

This is also usually computed using Monte-Carlo sampling and requires $O(N_S N_o N_\theta N_l)$ operations to evaluate one viewpoint, where N_S is the number of samples used.[2] As the dimensionality $N_o N_\theta N_l$ of the problem increases, the number of samples required to obtain a good approximation of Eq. (13) also increases, and it is usually the case that $N_S > N_o N_\theta N_l$.

Using Eq. (9) as an observation selection criterion with the approximation Eq. (11), the evaluation of one viewpoint takes $O(N_o^2 N_\theta^2 N_l)$ operations. The form of Eq. (9) lends itself to a further practical simplification: if the probability of a particular object class and pose pair is extremely low, then the terms in which the posterior probability for this pair appears contribute little to the sum and may be neglected. This causes the computation of Eq. (9) to get increasingly fast as the Bayesian inference engine converges toward a single winning hypothesis. This particular simplification is a consequence of the form of the chosen observation selection criterion. Indeed, expressing the utility of an observation as a sum of terms that perform pairwise comparisons naturally allows for some terms to be ignored when the probabilities they involve are below a certain threshold $p_{\min}$. This threshold can be adjusted empirically according to some reasonable trade-off between the cost of data acquisition and the cost of decision making.

5. Experiments

The active object recognition and pose estimation framework proposed in the previous sections is quite general and may be used in conjunction with a broad variety of feature extractors and appearance models. For the purpose of illustrating the proposed theory through experimentation, an appearance-based object representation based on the popular principal component analysis (PCA) (Murase and Nayar, 1995) was used for object representation. Using PCA, a low-dimensional feature space can be obtained from a small training set that adequately spans the range of possible appearances of a database of objects as seen under various viewing conditions. Note that any other feature extractor could have been used instead without modifying the overall framework.

A statistical appearance model can then be estimated from the projections of more training images (whose corresponding object class, pose, light source and viewpoint tuples are known) onto this feature space. The experiments presented in this paper make the common assumption (Arbel and Ferrie, 2001b; Borotschnig et al., 2000; Paletta and Pinz, 2000; Denzler and Brown, 2002); that the measurements are normally distributed given object class, pose, light source and viewpoint, i.e.

$$p(\mathbf{d} \mid o, \theta, l, \mathbf{v}) = \frac{\exp\left(-\frac{(\mathbf{d}-\mu(o,\theta,l,\mathbf{v}))^T \mathbf{C}(o,\theta,l,\mathbf{v})^{-1}(\mathbf{d}-\mu(o,\theta,l,\mathbf{v}))}{2}\right)}{\sqrt{(2\pi)^n |\mathbf{C}(o,\theta,l,\mathbf{v})|}}, \quad (14)$$

where $\mu(o, \theta, l, \mathbf{v})$ and $\mathbf{C}(o, \theta, l, \mathbf{v})$ respectively denote the mean vector and covariance matrix associated with object o in pose θ illuminated by light source l and seen from viewpoint $\mathbf{v}$, and n is the dimension of the feature space. The main advantage of using Gaussian distributions is their computational simplicity, which makes them useful in a broad variety of pattern recognition contexts. Should this assumption prove to be invalid, it could easily be changed to better fit true data acquisition conditions in a real world context without modifying the high level algorithms which are of interest in this work.

The remainder of this section focuses on active object recognition and pose estimation results pertaining to two case studies. The first case study examines active object recognition and pose estimation based on synthetic imagery obtained from 3D CAD models and a virtual camera with two degrees of freedom, with only one possible light source. Through experiments, it will be shown that in general, the proposed observation selection strategy requires fewer measurements to achieve recognition and pose estimation than a strategy based on a random walk. Also, the computational efficiency of the proposed strategy will be demonstrated by comparing it to a popular viewpoint selection strategy based on mutual information. The second case study explores the more difficult case of data acquired from a real imaging system with one degree of freedom and introduces illumination direction as a nuisance variable. This study emphasises the benefits of the proposed active strategy with respect to a passive, random strategy. The second case study also shows how the quality of the modeling and the experimental conditions can affect the accuracy of the object recognition and pose estimation results. This leads to a simple heuristic that improves the robustness of the proposed strategy to inaccuracies that result from poor modeling and/or poor experimental conditions.

Figure 3. Sample objects from the first case study.

5.1. *Case Study 1: Synthetic 3D Models*

The first case study was conducted with a database of 31 synthetic 3D models of aircraft.[3] Figure 3 shows sample rendered images of these objects. The motivation for using this database is that large amounts of images can easily be generated under a broad variety of conditions using common rendering packages. Also, the models are similar enough to pose a challenge to the recognition engine.

The problem considered was that of identifying an unknown object and estimating its pose under the illumination of a single possible light source. Clearly, in this context, no nuisance variables are modeled. The observer is a virtual camera with two degrees of freedom about a sphere, within which the object pose can vary according to two degrees of freedom (pan and tilt).

An object positioned with pan θ_1 and tilt θ_2 viewed from longitude 0 and latitude 0 can be thought of as the same object positioned with pan 0 and tilt 0 seen from longitude θ_1 and latitude θ_2. Clearly, pose and viewing position are defined on the same domain, i.e., the sphere $\mathbb{S}^2$, creating obvious equivalencies between different object pose and viewing position combinations. In particular,

$$p(\mathbf{d} \mid o, \theta, \mathbf{v}) = p(\mathbf{d} \mid o, \theta \oplus \mathbf{v}, \mathbf{0}) \qquad (15)$$

where $\mathbf{0}$ denotes the origin of the global reference frame and $\theta \oplus \mathbf{v}$ denotes the result of moving point θ by a geometric transformation equivalent to that required to move the origin of the reference frame to point $\mathbf{v}$. An appearance model is then adequately described by only the canonical conditional probability density function $p(\mathbf{d} \mid o, \theta, \mathbf{0})$, denoted by $p(\mathbf{d} \mid o, \theta)$ for the remainder of this section.

During a training phase, each model was rendered from 1380 randomly selected points of view about the viewing sphere. 1426 of the resulting images were used to construct a compact 3-dimensional feature space using PCA. Feature vectors were then computed for each of the training images by projection onto the resulting eigenspace. The set of possible object poses was discretised and reduced to 46 canonical poses, roughly uniformly spaced about the sphere. The space of possible viewing positions was discretised in the same fashion. A Gaussian distribution was then estimated for each class and pose pair (o, θ), using the sample feature vectors corresponding to viewing positions within 15 degrees of θ.

In a first set of experiments, the proposed observation selection strategy was compared to a random navigation strategy where no active selection of observations was employed. In the computation of the proposed observation selection criterion, all terms involving probabilities below the threshold $p_{\min} = 10^{-10}$ were neglected, as described in section 4. This set of experiments involved 50 recognition trials for each of the 31 synthetic models and each observation selection method. Convergence of the recognition process was considered attained when the entropy of the posterior distribution over object class and pose reached a confidence threshold of 0.1. The object recognition and pose estimation results are summarised in Table 1. It was found that upon convergence of the recognition process, both strategies achieved good accuracy in terms of correct classification rate and average pose estimation error, with no very significant differences. In terms of the number of measurements required to perform recognition, however, the proposed observation selection strategy was shown to be superior to a random navigation strategy. This is illustrated in Figure 4. These results show that the number of views required for recognition and pose estimation is almost consistently smaller for the proposed navigation strategy than for random navigation. The actual reduction in the number of observations required to achieve recognition (about 10%) is admittedly quite small for this case study, which is explained by the already small number of views required to solve the problem using random navigation. The next case study will demonstrate that more substantial improvements are obtained with more difficult recognition problems.

Table 1. Comparison of the accuracy of recognition and pose estimation results for the aircraft database, using the random and proposed observation selection strategies.

	Recognition rate	Average pose error
Random navigation	81%	1.84 degrees
Proposed strategy	83%	2.19 degrees

Similar experiments were then performed using an observation selection criterion based on mutual information (Denzler and Brown, 2002). The purpose of this was to compare the performance of the proposed criterion to a popular measure of informativeness which is known to yield good decisions and thus require fewer observations than a baseline random navigation strategy. Mutual information was evaluated using a Monte-Carlo sampling approach based on 1000 samples from the current posterior distribution $P(o, \theta | \mathbf{d}_t, \mathbf{v}_{t,\ldots,}\mathbf{d}_1, \mathbf{v}_1)$. This number was suggested in Denzler et al. (2002). Considering that there are 1426 hypotheses in this case study, this is a small number of samples. A better approximation of mutual information could have been obtained with more samples but this would have further increased the computational burden of the approach. Because of the latter computational burden (see Section 4), it was not feasible to conduct mutual information experiments on all 31 objects of the database, as each observation selection step took on the order of several minutes using the current implementation. Therefore, for the mutual information approach, experiments were only conducted on 14 of the 31 aircraft objects, with twenty trials per object instead of fifty, which nonetheless took several days.

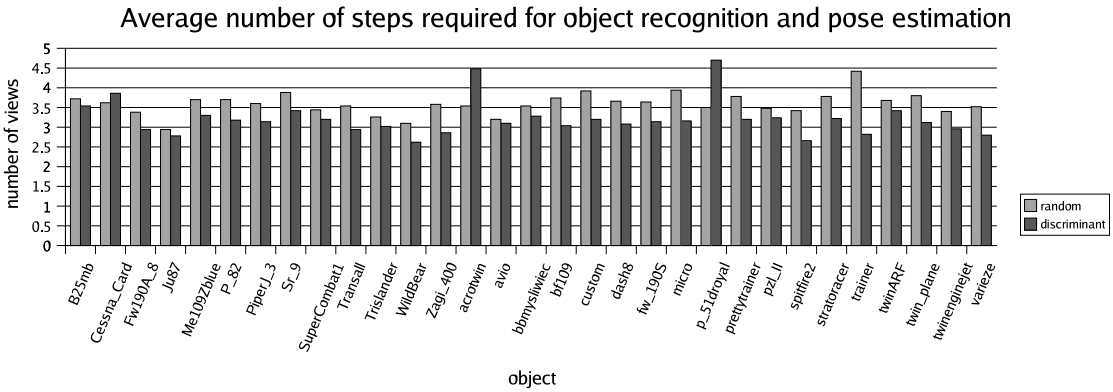


Figure 4. Comparison of the average number of steps required for recognition and pose estimation of the different objects in the aircraft database using both the random and proposed observation selection strategies.

The proposed strategy achieved similar results to mutual information, both in terms of accuracy and in terms of the number of views required for recognition. These results are summarised in Table 2. However, the proposed observation selection strategy is much less computationally expensive than mutual information. This is illustrated in Figure 5, where the progression of the amount of *time* required for decision making at each step is plotted for both strategies on a logarithmic scale. Notice that the amount of time required for the first decision is one order of magnitude lower with the proposed strategy than for the strategy based on mutual information. Furthermore, the time needed for decision making using the proposed strategy dramatically decreases as the recognition process progresses. To further reduce the time required for the first observation selection step, it remains possible to take a random step or use an off-line viewpoint selection method (e.g., Arbel and Ferrie, 2001a; Sipe and Casasent, 2002) for this first step.

Table 2. Performance comparison of the recognition and pose estimation results for 14 objects of the aircraft database, using the random, proposed and mutual information observation selection strategies.

	Recog. rate	Avg. pose error	Avg. views
Proposed strategy	82%	2.63 degrees	3.4 views
Mutual information	81%	0.43 degrees	3.3 views

5.2. *Case Study 2: Real Imagery*

The second case study considers the more general problem of object recognition and pose estimation under varying lighting conditions. In this context, object identity o and pose θ must be inferred from real imagery and the identity of the light source l is introduced as a nuisance variable. The study was conducted with a set of 13 objects which were custom-built with the purpose of rendering the recognition task difficult, and two light sources. Images of sample objects as seen from different points of view are shown in Figure 6 and one object is also shown as illuminated from the two possible light sources in Figure 7. As can be seen from Figure 6, not only do several objects share particular features, but some of them differ only in the way those features are arranged spatially.

An automated system was used to acquire training images of these objects. The setup consisted of a turntable with one degree of freedom in rotation and two degrees of freedom in translation, a camera placed at a fixed distance from the turntable, as well as an incandescent lamp on the right side of the turntable and a halogen lamp on the left side of the turntable. A black cloth was placed in the background to facilitate segmentation. Of course, provided with one of several segmentation algorithms in the literature, the object could be separated from the background. Segmentation was not the focus of this work, and the common assumption of black background (e.g. Murase and Nayar, 1995) was adopted. A photograph of the data acquisition setup is presented in Figure 8. Images of each object illuminated by each light source were automatically acquired for different object poses by rotating the turntable by 5 degree increments. This was repeated 18 times for each object and each light

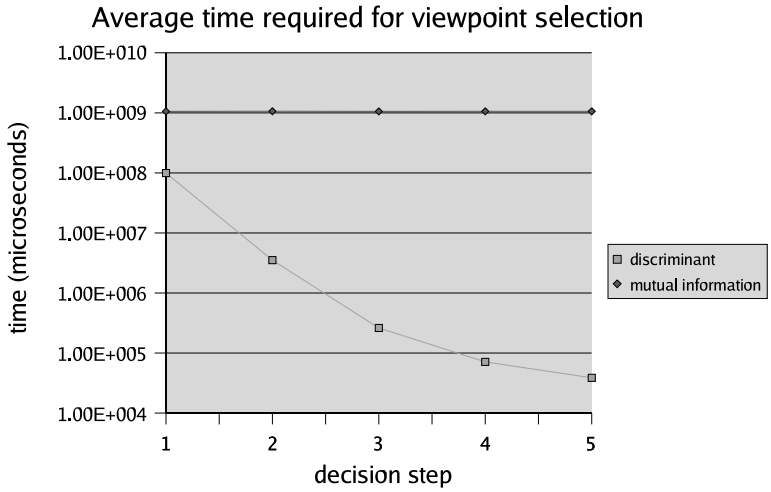


Figure 5. Comparison of the average time required for viewpoint selection as the recognition task progresses, using the mutual information and proposed observation selection strategies.

source, using the translational degrees of freedom of the turntable to impose slight variations in the position of the object with respect to the camera. As in the previous case study, there is a direct correspondence between viewpoint and object pose as viewing position is changed by rotating the object with respect to the camera about the same vertical axis. As before, this yields

$$p(\mathbf{d}|o,\theta,l,\mathbf{v}) = p(\mathbf{d}|o,\theta\oplus\mathbf{v},l,\mathbf{0}) \equiv p(\mathbf{d}|o,\theta\oplus\mathbf{v},l). \tag{16}$$

The space of possible object poses was reduced to 24 canonical poses, separated by 15 degree intervals. One image of each object in each canonical pose for each light source was removed from the training set, providing 624 test images for experimentation. Of the remaining training images, 936 were used to construct a 10-dimensional feature space through PCA. Feature vectors corresponding to the training images were then obtained by projecting these onto the resulting eigenspace. Gaussian distributions were then estimated for each object, pose and light source tuple (o, θ, l), using all the training images of object o with light source l corresponding to poses within 10 degrees of θ. In total, 89 images were used in the estimation of each Gaussian distribution. Once again, note that the proposed active recognition approach is independent of the particular statistical appearance model used and that any preferred model could be estimated in a similar manner.

5.2.1. Results Based on Measurements Sampled from the Appearance Distributions. For the sake of simplicity, the experimental framework described in this section introduces strong assumptions about the measurement process. An important step in validating the proposed active recognition strategy was to conduct an extensive series of fair experiments where these assumptions actually held. This was achieved through simulation of the data acquisition process by sampling feature vectors from the Gaussian distributions estimated during training. The observation selection strategy proposed in Section 4 (with parameter $p_{\min} = 10^{-5}$) was compared to a random navigation strategy. Note that comparison with a strategy based on mutual information was infeasible in this or any experiment pertaining to the second case study due to the computational intractability of this method.

In the first set of experiments, fifty active recognition trials were performed for each object and each observation selection method. It was considered that the system had converged when the entropy of the joint posterior distribution over object class and pose reached a confidence threshold of 0.1. For each trial, the pose of the object and the light source were selected at random. After convergence, it was found that both strategies yielded correct object classification rates very near 100%, with an average pose estimation error of 0 degree. The more interesting result is the number of observations required for convergence by each of the two methods, as presented in Figure 9.

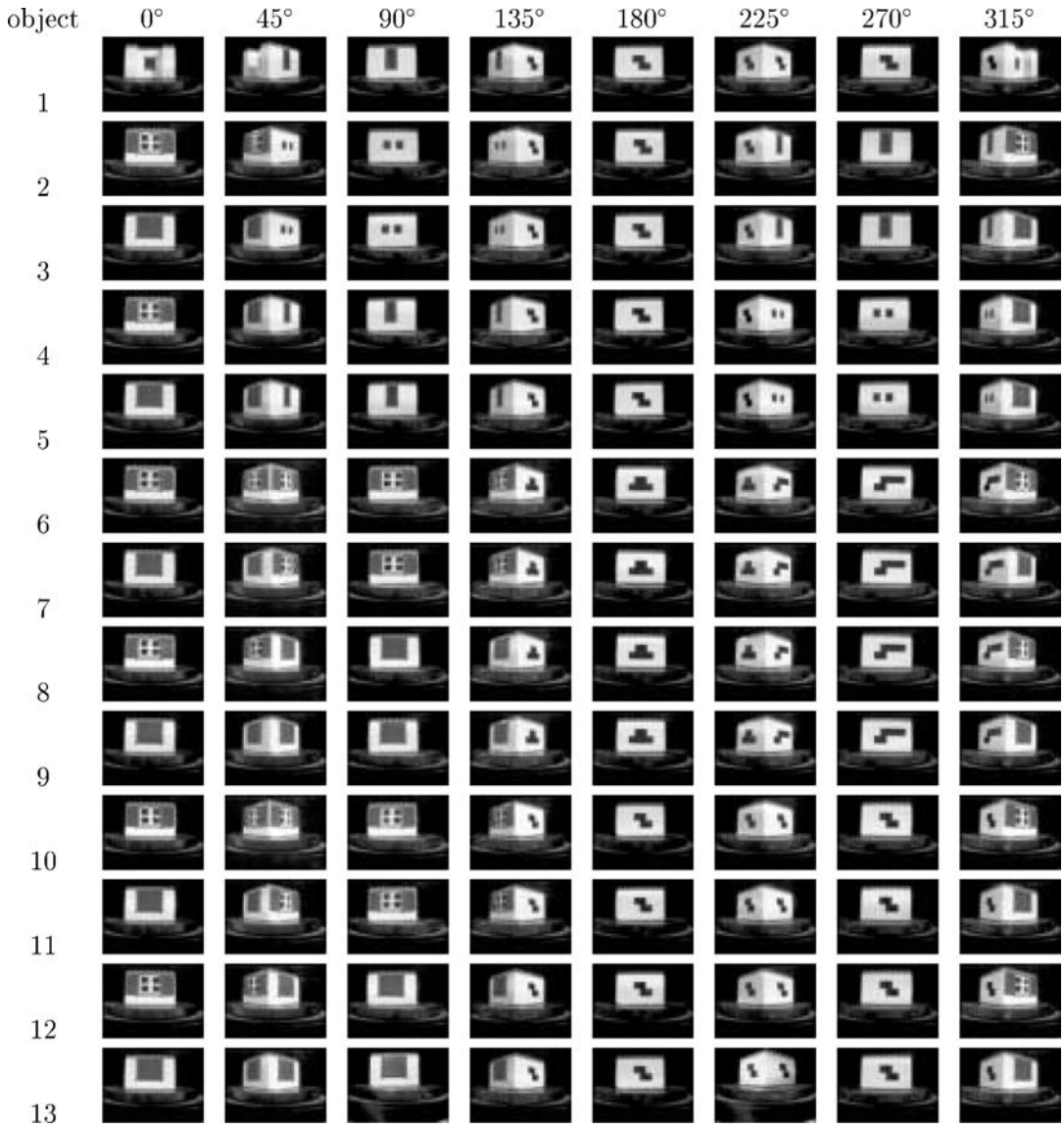


Figure 6. The 13 objects used for the second case study as seen from eight different points of view.

On average, the proposed navigation strategy achieved convergence in 5.29 data acquisition steps, whereas the random strategy took, on average, 7.73 steps. From these results, it is clear that, if the statistical appearance model is representative of the measurements, the observation selection strategy proposed in Section 4 significantly outperforms a random navigation strategy in terms of the number of views required to perform the task. A 30% reduction in the number of necessary observations (such as was obtained in this study) does not necessarily justify the use of even a moderately expensive active observation selection strategy such as this one when the cost of making the observations themselves is low. However, it is conjectured that the benefits (in terms of the number of observations required) of an active vision strategy increase with the difficulty of the problem (i.e., the degree of possible confusion inherent to the different hypotheses). Furthermore, there are cases when the acquisition of new data is very expensive or risky. For example, consider the case where the sensor is tied to an expensive robot evolving in an unknown and hostile

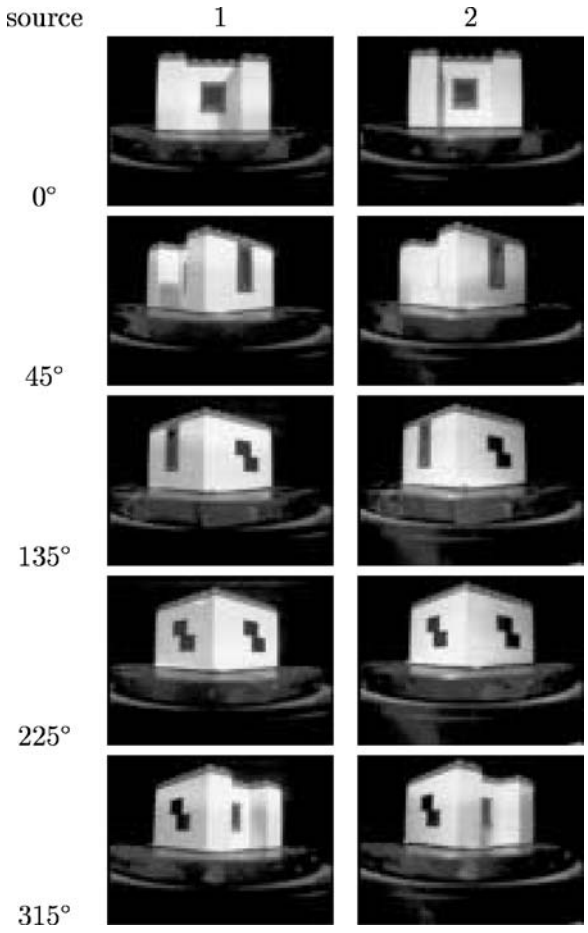


Figure 7. Sample views of object 1 illuminated from two different sources.

environment, where every movement of the sensor requires the robot to expose itself to potential danger.

5.2.2. Results Based on Real Image Data. The next step in validating the proposed observation selection strategy was to test it in a more realistic setting. Further experimentation was thus performed using the small pool of test data removed from the training set. As before with the simulated measurements, the proposed observation selection strategy (with $p_{\min} = 10^{-5}$) was compared with a random navigation strategy based on fifty active recognition trials for each object in the database, with the light source and object pose selected at random. In a first series of experiments, convergence was considered to have been attained when the entropy of the joint posterior distribution over object class and pose fell below a threshold of 0.1. The results are shown in Figure 10, where the average correct classification rate, pose estimation error and number of views required for convergence are plotted for each object in the database.

Predictably, the use of real imagery (as opposed to simulated measurements) introduced some error in the sequential recognition process, thereby yielding significantly lower recognition rates and higher pose estimation errors than in the simulated case, both for the random and proposed observation selection strategies. Despite these difficulties, Figure 10(c) shows that the proposed strategy required significantly fewer views to reach the desired level of confidence than the random navigation strategy. These results are investigated in more detail later in this section. Before this, the reasons for these difficulties must be fully understood.

The loss of accuracy introduced through the use real data is due to various violations of the assumptions that were made about the measurement process. In particular, significant error is introduced for the following reasons:

- *Misfit of the statistical appearance model to the test data*. The data do not necessarily fit the Gaussian model that was proposed for simplicity. Furthermore, the inference process is prone to over-fitting problems as relatively few samples were used to estimate the parameters of these Gaussian distributions (89 samples for 65 parameters).
- *Violation of the independence assumption*. The recursive data fusion methodology presented in section 3 relies on the assumption that subsequent measurements are independent of one another given the state of the world under observation and the observation parameters, i.e., that the variations in the measurements of a given object in a given canonical pose under a given light source are entirely due to independently and identically distributed noise. However, real vision systems are also subject to noise that is not independently and identically distributed, such as variations in ambient lighting that occur gradually as the day goes by. Such latent sources of noise act as unmodeled nuisance variables in the system, and cause the independence assumption to be violated.

According to Figure 10, the proposed observation selection strategy was affected by these problems to a much greater extent than the random navigation strategy. In fact, the correct object classification rate after convergence is extremely poor for some objects (in particular, objects 4, 6 and 10), as well as some of the pose

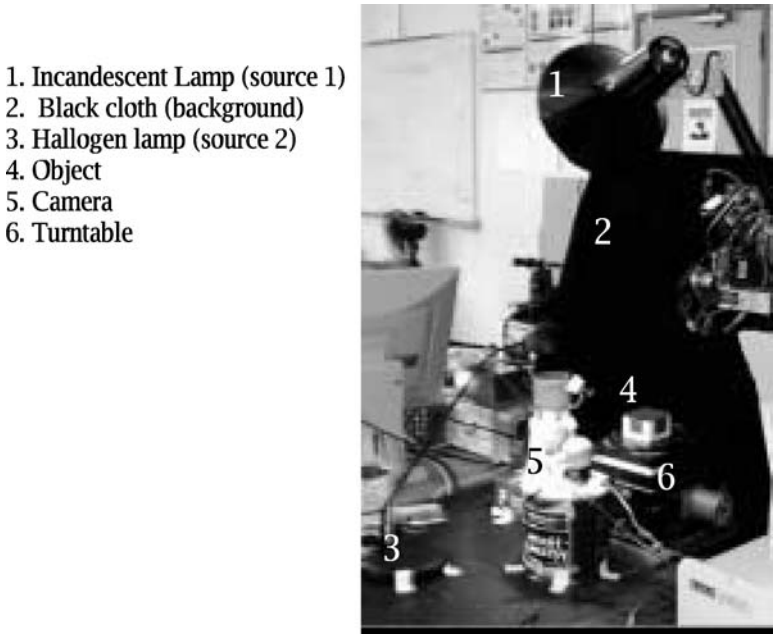


Figure 8. The image acquisition setup used to acquire data for training and experiments.

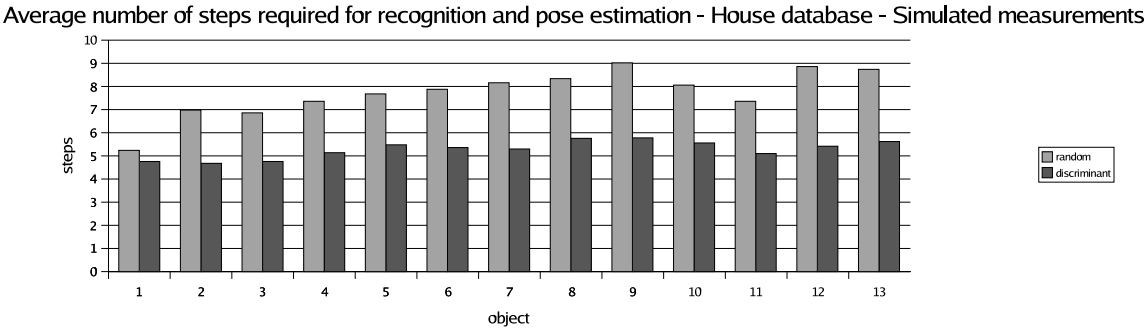
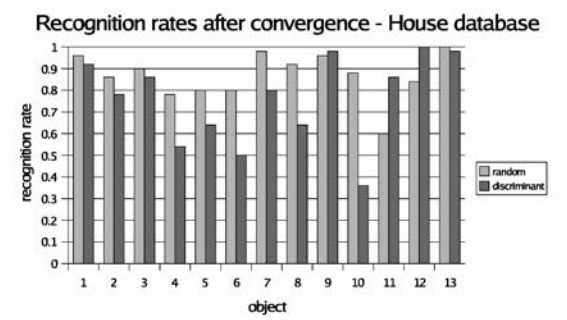


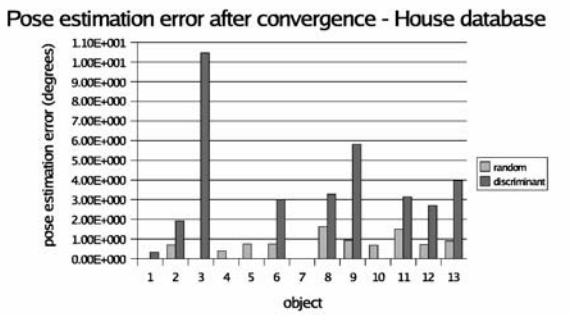
Figure 9. Average number of steps required for recognition and pose estimation of the different house objects, based on simulated measurements.

estimates (for example, in the case of object 3). Under the described experimental conditions, two important factors explain this. The first explanation results from the bootstrapping effect that is induced by the use of an active observation selection strategy through the use of the statistical appearance model for both inference and decision making, as opposed to only inference in the random case. Misfit between the model and the test data may cause the observation selection algorithm to make bad decisions, based on erroneous information. The second explanation stems from the fact that, in contrast with a random navigation strategy, the proposed observation selection strategy is prone to repeatedly select the same viewpoint if it happens to be particularly useful for discrimination. Repeating observations can significantly aggravate the problems due to incorrect modeling assumptions.

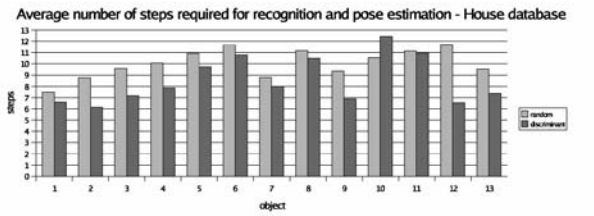
While there is little sign of this problem when the appearance model is good (see Section 5.2.1), the error inducing pattern caused by repeating observations under conditions where the chosen statistical appearance model was poor was easy to observe empirically. In particular, consider the case of object 10, which was only classified correctly in 36% of cases, often being confused with object 12 in the same pose. This behaviour was investigated in more detail, and it was found that the proposed observation selection strategy consistently led the sensor to observe the 90 degree



(a) Average recognition rate for each object in the database.



(b) Average pose estimation error for each object in the database.



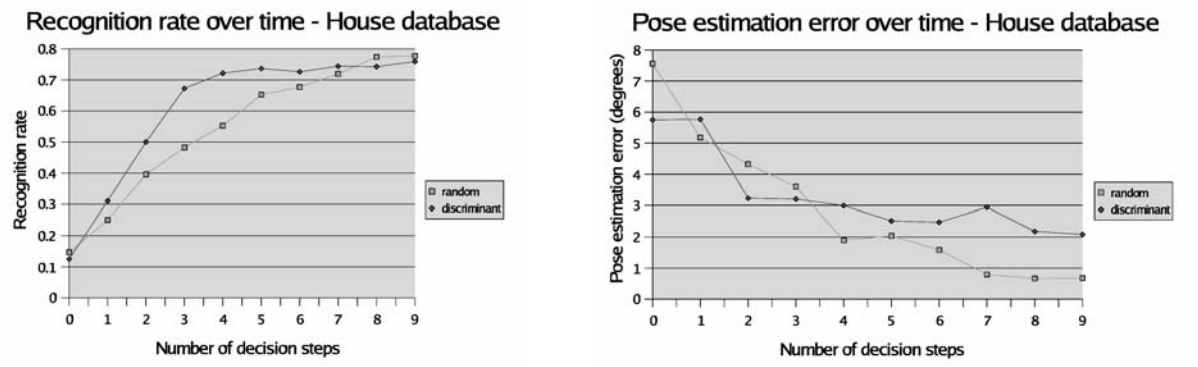
(c) Average number of steps required for recognition and pose estimation of each object in the database.

Figure 10. Active recognition results after convergence of the sequential recognition process, based on real data from the house database.

canonical view of object 10 (see Figure 6). From a human perspective, this appears to be a good choice as this view of object 10 significantly differs from the analogous view of object 12 from a qualitative point of view. However, deeper analysis revealed a significant flaw in the statistical appearance model for this case. Indeed, according to the model, the difference between the means of the two classes for this particular pose (regardless of the light source) is very small in comparison to the difference in variance. Further investigation has shown that, in fact, the Gaussian is a poor model in this case. The inaccurate results in these conditions are entirely due to poor modeling. Recall that, as shown in the previous set of experiments (see Section 5.2.1), the proposed active observation selection strategy works very well when the underlying appearance model fits the data.

Despite the accuracy problems due to poor modeling, the proposed strategy required a significantly smaller number of views to reach the desired level of confidence than the random navigation strategy, as illustrated in Figure 10(c). This means that the proposed observation selection strategy does indeed make intelligent decisions (i.e. ones that will reduce the number of observations needed to perform recognition) *with respect to the information available to it*. If this information happens to be erroneous, it will reflect in the accuracy of the recognition results after convergence, as the evaluation of the stopping criterion (i.e. entropy) is performed on distributions calculated from erroneous *a priori* information as though it were valid.

A second set of experiments was performed under the same conditions, this time stopping the data acquisition after ten observations were made in order to analyse the evolution of accuracy over time. The average recognition rate and pose estimation error were recorded at each step for both the random and proposed observation selection strategies. The results are presented in Figure 11. As shown in Figure 11(a), during the first few steps of the active recognition process, the average recognition rate obtained with the proposed observation selection strategy grows at a much faster pace than that obtained with the random walk approach. However, after a few data acquisition steps, the average recognition rate of the proposed strategy appears to reach a ceiling (around 75% correct



(a) Average recognition rates over time. (b) Average pose estimation errors over time.

Figure 11. Evolution of the average recognition rates and pose estimation error over time for the random and proposed navigation strategies, based on real image data.

classification), while the random strategy still makes progress over time. This result strongly suggests that the performance of the proposed active recognition strategy is indeed affected by an over-fitting problem (due to bad modeling). At a very coarse level, it appears that the proposed model is sufficiently accurate to ensure that the proposed observation selection strategy makes good decisions, which explains its relatively good behaviour during the first few data acquisition steps. At finer levels of detail, however, the model is not accurate enough to ensure that the active observation strategy will make decisions that will truly disambiguate inherently similar, competing hypotheses. Rather, too much reliance on the poor model causes the observation selection strategy to expect useful information where this information is not present. The results pertaining to the evolution of the pose estimation error over time support this hypothesis. The next section proposes a simple remedy to this problem.

5.2.3. Increasing Robustness to Inaccuracies in Modeling Assumptions. Arguably, the obvious remedy to the inaccuracies due to poor modeling would be to make more accurate assumptions, possibly based on a more substantial pool of training data. Although it is true that the modeling process could be improved upon and the underlying assumptions modified, it is only fair to expect that any realistic experimental conditions will also violate the proposed assumptions to some (although perhaps lesser) extent. Ideally, one would like the proposed solution to the active recognition problem to be somewhat robust to such inaccuracies, so that it may be usefully implemented in a real vision system. As discussed earlier, it appears that the baseline random navigation strategy is less sensitive to the quality of the modeling and experimental setup than the proposed active vision strategy. This suggests a certain trade-off between the number of observations needed for recognition and robustness to poor modeling. Conceivably, this could be achieved by judiciously alternating the use of the two methods, depending on what is known about the quality of the modeling assumptions. Instead, this section proposes a heuristic strategy which attempts to exploit the non-repetitiveness characteristic inherent to random navigation which contributes to its robustness.

As stated earlier, an important source of error in the context of the proposed active recognition strategy is repeating observations. These aggravate the problems

Table 3. Comparison of the results obtained with and without the non-repeating navigation constraint for the random and proposed view-point selection approaches.

	Recog. rate	Avg. pose error	Avg. views
Random	87%	0.69 degrees	10.06
Proposed strategy	76%	2.66 degrees	8.54
Random non-repeating	93%	0.49 degrees	9.36
Proposed strategy non-repeating	94%	1.71 degrees	6.85

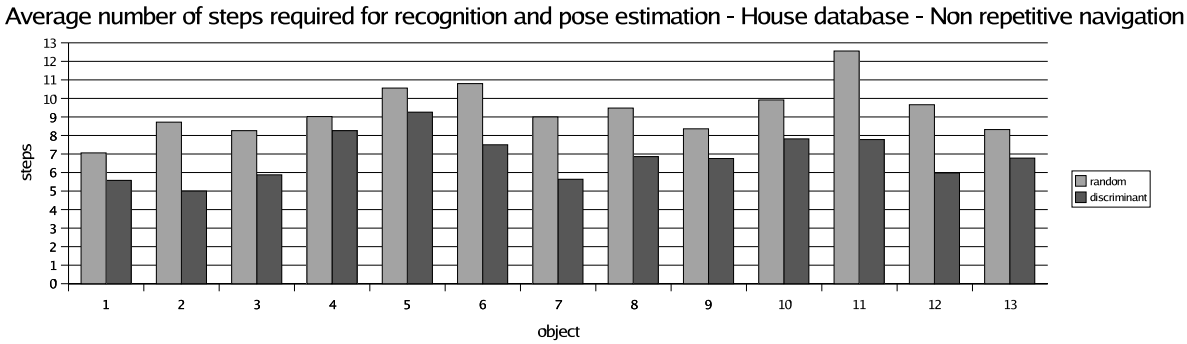


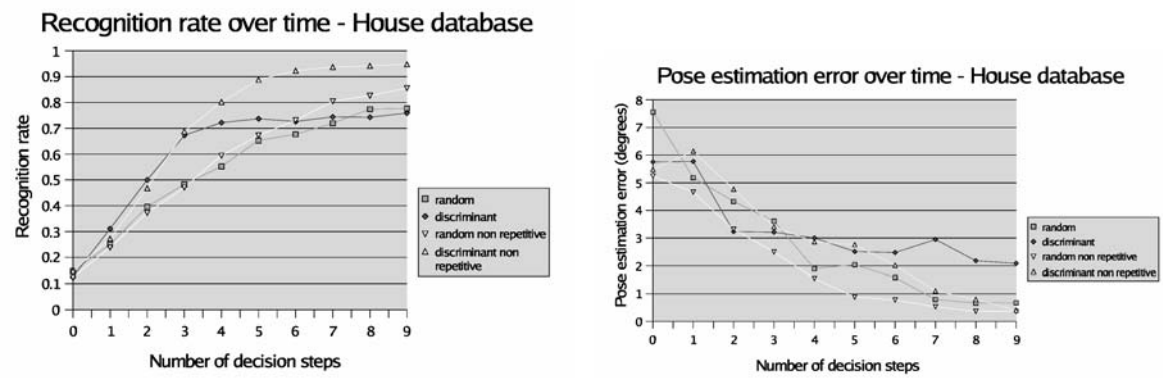
Figure 12. Comparison of the average number of steps required for recognition and pose estimation of the house-like objects of the second case study using both the random and proposed observation selection strategies.

caused by the violation of the assumptions underlying the recognition process. This effect was also reported by Borotschnig et al. (2000) in the case of a different observation selection criterion. In the context of viewpoint selection, the authors propose the use of a heuristic mask to avoid revisiting locations where measurements were recently acquired, a method akin to the well-known Tabu search techniques documented in the operations research and artificial intelligence literature (Glover and Laguna, 1998; Russel and Nurvig, 1995). This enforces global exploration of the scene, thereby making active observation selection approaches more robust to the combined effects of over-fitting, non-markovianity and bootstrapping. In the following sets of experiments, a similar heuristic is employed; that is, the sensor is forbidden to visit a given viewing position more than once. The proposed active observation selection strategy was thus modified to select the best option among the ones that were not already explored. For the sake of fairness, the modified strategy is compared to a non-repeating random navigation strategy that uses the same heuristic.

As before, fifty recognition trials were performed for each object, with the light source and object pose selected at random, and stopping the data acquisition when the entropy of the joint posterior distribution over object class and pose fell below 0.1. Table 3 and Figure 12 summarise the results. From Table 3, it is clear that introducing the non-repeating navigation constraint has improved the accuracy of the recognition results, especially in the case of the proposed navigation strategy. The average rate of correct classification obtained with the non-repeating version of the proposed observation selection strategy compares with that obtained with the non-repeating random navigation strategy. Nonetheless, the results pertaining to pose estimation error show that, despite a major improvement in accuracy with the introduction of the non-repeating navigation constraint, the proposed observation selection strategy is still slightly less accurate than its random counterpart. This is due to the remaining effects of the combination of model misfit and bootstrapping. Note, however, that the average pose estimation error of 1.71 degrees obtained with the proposed navigation strategy (compared to 0.49 degrees in the random case), is still quite low and sufficiently accurate for many applications where pose estimation is regarded as a secondary goal or by-product of object recognition.

The slight degradation in the accuracy of the pose estimates in the case of the proposed observation selection strategy is largely compensated by the lower cost implied by the acquisition of measurements. As shown in Figure 12, the number of views required for recognition and pose estimation of the different objects is consistently and significantly lower in the case of the proposed navigation strategy (on average, 6.85 views) than in the random case (9.36 views on average). These results show that the difficulties encountered by the proposed navigation strategy due to inaccurate modeling can largely be overcome by introducing a simple non-repeating navigation heuristic, thereby illustrating the feasibility and practicality of its implementation in a real world environment.

Again, further experiments were performed where the data acquisition was stopped after 10 views while the average recognition rates and pose estimation errors were recorded at every step for the non-repeating versions of the random and proposed observation selection strategies. The results are shown in Figure 13, where



(a) Evolution of the recognition rates over time for the different navigation strategies.

(b) Evolution of the pose estimation error over time for the different navigation strategies.

Figure 13. A comparison between the repeating and non-repeating versions of the random and proposed navigation strategies based on the evolution of recognition accuracy over time, using real data.

they are plotted against the analogous results obtained without the non-repeating navigation heuristic. The practical benefits of the heuristic are very apparent in these results, especially in the case of the proposed observation selection strategy. The non-repeating version of the proposed strategy behaves very much like the original one for the first few steps, with a fast increase in the rate of correct classification. However, as time goes by, the non-repeating version of the proposed strategy does not stop making progress as the original one did and outperforms all the other illustrated strategies. The improvement can also be seen in the evolution of the pose estimation error, where the non-repeating version of the proposed observation selection strategy exhibits much better behaviour than its original counterpart. This shows that introducing a non-repeating navigation constraint largely overcomes the strong over-fitting problems that were observed in previous experiments, thereby allowing an active recognition system to benefit from the power of the observation selection strategy proposed in Section 4 without a significant loss of accuracy.

6. Conclusion

This paper presented a new observation selection strategy for active recognition problems that allows for competing hypotheses to be effectively disambiguated and is an efficient alternative to popular techniques that maximise mutual information or average loss of entropy. The method was applied to two different and difficult object recognition and pose estimation problems involving synthetic and real data, respectively. Experiments showed that an object recognition and pose estimation system based on the inference and observation selection schemes presented in this paper was effective. Furthermore, the proposed observation selection strategy was shown to be much quicker than a strategy based on mutual information, and to require fewer measurements on average than a random navigation strategy. On its own, the proposed observation selection strategy does not compromise the accuracy of the results provided the assumptions underlying the chosen appearance model fit the measurements. Moreover, experimentation demonstrated that, should these assumptions be violated, loss of accuracy is largely overcome by introducing a non-repeating navigation constraint to increase the robustness of the proposed algorithm, thereby leading to an active vision system that is efficient without compromising the accuracy of the recognition results. Conceivably, the new approach could be combined with instance-based learning techniques to further accelerate the viewpoint selection process. That is, the system could select a viewpoint by consulting a map from posterior distributions or sequences of previous actions to optimal viewpoints constructed off-line as in (Paletta et al., 2000). The suggested strategy would then provide a means of drastically reducing the cost of the training phase.

Notes

1. Note that the computational cost of this modeling phase is, once again, not unique to the particular approach described in this paper. In particular, many viewpoint selection methods based on dynamic programming or reinforcement learning require this modeling stage *in addition* to the prohibitive computation of the viewpoint selection policy.
2. Note that in most cases, mutual information is impossible to compute analytically, which is why approximate methods such as Monte-Carlo sampling are used. An alternative would be some form of numerical quadrature, which is also computationally expensive as well as difficult to implement, explaining the popularity of the Monte-Carlo approach.
3. Radio control—computer aided design gallery. http://www.rccad.com/Gallery–Classic8.htm.

References

Aloimonos, J., Weiss, I., and Bandyopadhyay, A. 1988. Active vision. *International Journal of Computer Vision,* 1(4):333–356.

Arbel, T. and Ferrie, F.P. 2001. Entropy-based gaze planning. *Image and Vision Computing,* 19:779–786.

Arbel, T. and Ferrie, F.P. 2001. On the sequential accumulation of evidence. *International Journal of Computer Vision,* 43(3):205–230.

Arbel, T. and Ferrie, F.P. 2002. Interactive visual dialog. *Image and Vision Computing,* 20:639–646.

Bajcsy, R. 1988. Active perception. *Proceedings of the IEEE,* 76(8):966–1005.

Borotschnig, H., Paletta, L., Prantl, M., and Pinz, A. 2000. Appearance-based active object recognition. *Image and Vision Computing,* 18:715–727.

Callari, F.G. and Ferrie, F.P. 2001. Active object recognition: Looking for differences. *International Journal of Computer Vision,* 43(3):189–204.

Chen, J.-H. and Chen, C.-S. 2004. Object recognition based on image sequences by using inter-feature-line consistencies. *Pattern Recognition,* 37:1913–1923.

Cover, T.M. and Thomas, J.A. 1991. *Elements of Information Theory,* Wiley.

Darrell, T. and Pentland, A. 1995. Active gesture recognition using learned visual attention. In *Neural Information Processing Systems*, vol. 8, pp. 858–864.

Denzler, J. and Brown, C.M. 2002. Information theoretic sensor data selection for active object recognition and state estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):145–157.

Dickinson, S.J., Christensen, H.I., Tsotsos, J.K., and Olofsson, G. 1997. Active object recognition integrating attention and viewpoint control. *Computer Vision and Image Understanding*, 67(3):239–260.

Ertin, E. and Priddy, K.L. 2002. Reinforcement learning and design of nonparametric sequential decision networks. In *Proceedings of the SPIE Conference on Applications and Science of Computational Intelligence IV*.

Geman, D. and Jedynak, B. 2001. Model-based classification trees. *IEEE Transactions on Information Theory*, 47(3):1075–1082.

Glover, F. and Laguna, M. 1998. *Tabu Search*. Kluwer Academic Publishers.

Goldberger, J., Gordon, S., and Greenspan, H. 2003. An efficient image similarity measure based on approximations of KL-divergence between two Gaussian mixtures. In *Proceedings of the 9th International Conference on Computer Vision*, Nice, France, pp. 487–493.

Gremban, K.D. and Ikeuchi, K. 1994. Planning multiple observations for object recognition. *International Journal of Computer Vision*, 12(2):137–172.

Herbin, S. 1996. Recognizing 3D objects by generating random actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, San Francisco, CA, pp. 35–40,

Kaelbling, L.P., Littman, M.L., and Cassandra, A.R. 1998. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101:99–134.

Kovačič, S., Leonardis, A., and Pernuš, F. 1998. Planning sequences of views for 3-D object recognition and pose determination. *Pattern Recognition*, 31(10):1407–1417.

Laporte, C. 2004. A fast discriminant approach to active Bayesian visual recognition and pose estimation. Master's thesis, McGill University, Montreal, Quebec, Canada.

Laporte, C., Brooks, R., and Arbel, T. 2004. A fast discriminant approach to active object recognition and pose estimation. In *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge, U.K, vol. 3, pp. 91–94,

MacKay, D.J.C. 1992. Bayesian interpolation. *Neural Computation*, 4(3):415–447.

MacKay, D.J.C. 1992. Information-based objective functions for active data selection. *Neural Computation*, 4(4):589–603.

Murase, H. and Nayar, S.K. 1995. Visual learning and recognition of 3-D objects from appearance. *International Journal of Computer Vision*, 14:5–24.

Paletta, L. and Pinz, A. 2000. Active object recognition by view integration and reinforcement learning. *Robotics and Autonomous Systems*, 31:71–86.

Paletta, L., Prantl, M., and Pinz, A. 2000. Learning temporal context in active object recognition using Bayesian analysis. In *Proceedings of the 15th International Conference on Pattern Recognition*, Barcelona, Spain, pp. 695–699.

Russel, S. and Norvig, P. 1995. *Artificial Intelligence: A Modern Approach*, Prentice-Hall.

Schiele, B. and Crowley, J.L. 1998. Transinformation for active object recognition. In *Proceedings of the International Conference on Computer Vision*, Bombay, India, pp. 249–254.

Seibert, M. and Waxman, A.M. 1992. Adaptive 3-D object recognition from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):107–124.

Sipe, M.A. and Casasent, D. 2002. Feature space trajectory methods for active computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(12):1634–1643.

Sutton, R.S. and Barto, A.G. 1998. *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA.

Vasconcelos, N. 2001. On the complexity of probabilistic image retrieval. In *Proceedings of the 8th International Conference on Computer Vision*. Vancouver, Canada.. pp. 400–407.

Zhou, X.S., Comaniciu, D., and Krishnan, A. 2003. Conditional feature sensitivity: A unifying view on active recognition and feature selection. In *Proceedings of the 9th International Conference on Computer Vision*, Nice, France, pp. 1502–1509.