A fast discriminant approach to active object recognition and pose estimation

Catherine Laporte, Rupert Brooks, Tal Arbel
McGill University
Centre for Intelligent Machines
Montreal, Quebec, Canada
cathy@cim.mcgill.ca, rbrook@cim.mcgill.ca, arbel@cim.mcgill.ca

Abstract

This paper presents a new criterion for viewpoint selection in the context of active Bayesian object recognition and pose estimation. Recognition is performed by probabilistically fusing successive observations with the current belief state of the system. Based on the current belief state, the next viewpoint is chosen to maximize the expected discriminability of the current competing hypotheses. Experiments on a difficult database of aircraft models show that this approach achieves comparable recognition performance to the widely used information theoretic approaches at a much lower computational cost.

1. Introduction

Object recognition techniques have been the focus of a large amount of literature in the past decade because of their direct application to real-world problems. The problem of object recognition in the presence of high uncertainty led to the development of systems that perform recognition using a sequence of observations from different points of view. It has been shown that the number of views required to recognize objects with a given level of certainty can be reduced by actively selecting the viewpoints from which the different images are gathered according to some measure of their discriminative power.

The performance of an active object recognition system depends on two major components: an inference component which fuses the evidence accumulated from successive observations, and an observation selection component which chooses the parameter settings for the next observation. In this work, probabilistic inference is employed as it is a particularly effective method for evidence fusion. Previous work has explored mainly two classes of viewpoint selection algorithms: off-line strategies based on a precomputed set of characteristic views[11, 2, 12] and on-line strategies that attempt to maximize information gain given the current belief state of the system[3, 10, 5, 13]. On-line algorithms have generally used information theoretic measures of discriminability to choose the next viewpoint. They tend to perform better in terms of the number of views required to achieve recognition; however, they also tend to be much more computationally expensive than off-line ones. The computational complexity of on-line gaze planning algorithms usually increases drastically with the number of possible labels.

This article presents a new viewpoint selection criterion that reduces the cost of decision making. The approach seeks to choose viewpoints that maximize the separability of competing hypotheses. In this work, the primary goal is not only object identification, but also pose estimation. The dimensionality of this problem is such that traditional on-line gaze planning techniques are prohibitively expensive.

The paper is structured as follows: section 2 outlines the Bayesian inference scheme used for sequential object recognition and pose estimation. Section 3 describes previous approaches to viewpoint selection, and section 4 sets out the proposed criterion. Section 5 validates the proposed approach through rigorous experiments on a large set of similar objects. The proposed approach is shown to outperform a random viewpoint selection strategy and have comparable performance to that of the traditional information theoretic criteria while being much less costly.

2. Sequential Bayesian recognition

Consider a database of objects \(O_i, i \in \{1, \ldots, n\}\) and a mobile camera facing an object whose class and pose are to be identified. Let the camera measurement be parameterized by a feature vector \(d\), which depends on the identity \(O_i\) of the object, its pose \(\theta\) and the viewing position \(v\). Under uncertainty, this relationship can be represented through a probability density function \(p(d|O_i, \theta, v), i = 1, \ldots, n; \theta \in S^2, v \in V\) (where \(S^2\) is the surface of a unit sphere and \(V\) denotes the set of possible camera viewpoints) whose parameters are assumed to be learned or modeled off-line through some training procedure.
Given a known viewing position \( \mathbf{v} \), and a prior distribution \( P(o_i, \theta) \) over object class and pose, the probability of the object belonging to each class-pose pair given a measurement \( \mathbf{d} \) is computed using Bayes’ rule:

\[
P(o_i, \theta | \mathbf{d}, \mathbf{v}) = \frac{1}{K} p(\mathbf{d} | o_i, \theta, \mathbf{v}) P(o_i, \theta), \tag{1}
\]

where \( K \) is a normalizing constant.

Following this assumption, the conditional dependence of multiple observations of a particular object given the viewing positions from which they are taken leads to the recursive observation fusion scheme described in [10]:

\[
P(o_i, \theta | \mathbf{d}_1, \mathbf{v}_1, \ldots, \mathbf{d}_t, \mathbf{v}_t) \propto p(\mathbf{d}_t | o_i, \theta, \mathbf{v}_t) P(o_i, \theta | \mathbf{d}_1, \mathbf{v}_1, \ldots, \mathbf{d}_{t-1}, \mathbf{v}_{t-1}), \tag{2}
\]

where \( \mathbf{v}_t \) and \( \mathbf{d}_t \) are the viewing position and observation at time step \( t \). As more observations of an object are taken, a sequential recognition engine based on this inference model exploits the information provided by the appearance of the object and, more importantly, by its spatial structure.

### 3. Viewpoint selection

The active object recognition and pose estimation problem can be defined as that of finding the viewpoint selection strategy that minimizes the number of observations required to perform recognition and pose estimation of an unknown object with a particular level of confidence. This strategy is dependent on the relationship between camera observations, object class, object pose and camera parameters (i.e. viewing position).

The theoretically optimal solution to this problem requires analysis of all possible sequences of future decisions. Such global solutions were suggested in the context of active gesture recognition [4], active object recognition [9] as well as the similar framework of decision trees [7]. The main drawback of these solutions is their computational complexity, which makes them prohibitively expensive on large domains.

Instead, several approaches for active recognition use a myopic decision policy; that is, they attempt to maximize the information gain of the action taken at every step. Using Shannon entropy as a measure of ambiguity, this corresponds to maximizing the expected mutual information between viewing position and the variables of interest. For the problem of joint object recognition and pose estimation, this leads to selecting the viewpoint at time \( t + 1 \) such that

\[
\mathbf{v}_{t+1} = \arg\max_{\mathbf{v}_{t+1}} \mathbb{E}[H(O, \Theta | \mathbf{d}_1, \mathbf{v}_1, \ldots, \mathbf{d}_t, \mathbf{v}_t) - H(O, \Theta | \mathbf{d}_1, \mathbf{v}_1, \ldots, \mathbf{d}_{t+1}, \mathbf{v}_{t+1})], \tag{3}
\]

where \( H(\cdot) \) denotes the entropy of a random variable.

Myopic gaze planning strategies based on similar measures have been shown to work quite well [3, 10, 5]. Despite the conceptual simplicity of this local measure of viewpoint utility, its evaluation suffers from computational complexity when the dimensionality of the active recognition problem is large. In [10], this issue is addressed by transferring the computational load to an extra training phase that constructs a map from past sequences of viewing positions to the best next viewing positions according to (3). On-line, a nearest neighbour approach is used to find the closest trained sequence of actions to the current history of the system and apply the corresponding camera displacement.

The problem addressed in this paper includes object pose estimation in addition to object identification, which increases the complexity of evaluating (3) by a factor of the order of the number of possible object poses. Instance based learning such as in [10] then becomes prohibitively expensive, even though it is done off-line.

### 4. Proposed Approach

Rather than attempt to accelerate the computation of (3), this paper presents an efficient, alternative, viewpoint selection criterion that quantifies the expected discriminative power of a viewpoint; that is, its ability to present new visual information in such a way as to disambiguate competing hypotheses. The approach is inspired by Fisher’s Linear Discriminant Analysis (LDA), a technique commonly used in pattern classification to reduce the dimensionality of a set of data points while retaining the ability to classify them.

LDA is intended for use on a multidimensional set of data points divided into two different classes. It finds the projection of the data onto a line which maximizes the separability of the distributions of the projected data points in each class. Specifically, LDA yields the vector \( \mathbf{w} \) such that:

\[
\mathbf{w} = \arg\max_{\mathbf{w}} \frac{(\mu_1(\mathbf{w}) - \mu_2(\mathbf{w}))^2}{\sigma_1^2(\mathbf{w}) + \sigma_2^2(\mathbf{w})}, \tag{4}
\]

where \( \mu_1(\mathbf{w}) \) and \( \mu_2(\mathbf{w}) \) are the means and \( \sigma_1^2(\mathbf{w}) \) and \( \sigma_2^2(\mathbf{w}) \) are the variances of the data points in each class, as projected by \( \mathbf{w} \) [6].

Under certain conditions, LDA can be interpreted as a method for choosing an optimal viewing direction. Consider the data to be 2-D points, classified into two groups. The LDA solution is the line along which the projections of the two classes are maximally separated. For an imaginary observer at infinity, the optimal viewing direction to classify a new data point is perpendicular to the LDA line.

However, in the active recognition problem, there are several differences which preclude the direct application of LDA. Most importantly, there are usually more than two classes to separate. Unfortunately, the natural generalization of LDA to more than two classes becomes as com-
putationally complex as (3). Furthermore, a viewpoint is sought rather than a linear operator, and the distributions were learned during training rather than being computed.

Therefore, a measure is sought that maximizes the separability of each pair of classes, while incorporating the relative likelihood of encountering each pair. The following sum, weighted by the probabilities of each class, is proposed as a good criterion:

$$\mathbf{v}_{t+1}^{*} = \arg\max_{\mathbf{v}_{t+1}} \sum_{i=1}^{n_{o}} \sum_{j=1}^{n_{v}} P(o_i, \theta_j | \mathbf{d}_1, \mathbf{v}_1, \ldots, \mathbf{d}_i, \mathbf{v}_i)$$

$$\sum_{k=1}^{n_{o}} \sum_{l=1}^{n_{v}} P(o_k, \theta_l | \mathbf{d}_1, \mathbf{v}_1, \ldots, \mathbf{d}_l, \mathbf{v}_l)$$

$$M(o_i, \theta_j, o_k, \theta_l, \mathbf{v}) = \frac{1}{M(o_i, \theta_j, o_k, \theta_l, \mathbf{v}_{t+1})}. \quad (5)$$

Here, $$M(o_i, \theta_j, o_k, \theta_l, \mathbf{v})$$ is a measure of separability between the expected observations for object $$o_i$$ in pose $$\theta_j$$ and object $$o_k$$ in pose $$\theta_l$$ given by:

$$M(o_i, \theta_j, o_k, \theta_l, \mathbf{v}) = \left( \mathbf{\bar{d}}(i, j, \mathbf{v}) - \mathbf{\bar{d}}(k, l, \mathbf{v}) \right)^{T}$$

$$\left( \mathbf{C}_d(i, j, \mathbf{v}) + \mathbf{C}_d(k, l, \mathbf{v}) \right)^{-1} \times$$

$$\left( \mathbf{\bar{d}}(i, j, \mathbf{v}) - \mathbf{\bar{d}}(k, l, \mathbf{v}) \right), \quad (6)$$

where $$\mathbf{\bar{d}}(i, j, \mathbf{v})$$ and $$\mathbf{C}_d(i, j, \mathbf{v})$$ are the mean and covariance matrix, respectively, of the expected observation of object $$o_i$$ in pose $$\theta_j$$ from viewpoint $$\mathbf{v}$$. It is important to note that this measure depends, as does LDA, on the mean and covariance matrix being a good representation of the distribution of the observations.

Observe that the quantities $$M(o_i, \theta_j, o_k, \theta_l, \mathbf{v})$$ may be entirely computed off-line. Thus, during the run, each viewpoint evaluation only requires $$n_o^2 n_v^2$$ additions and $$2n_o^2 n_v^2$$ multiplications. In practice, a further simplification is obtained by observing that if the probability of either hypothesis, $$P(o_i, \theta_j | \mathbf{d}_i, \mathbf{v}_i)$$ or $$P(o_k, \theta_l | \mathbf{d}_k, \mathbf{v}_k)$$, is extremely low (say, $$10^{-10}$$), then the term on the right does not affect the sum appreciably and may be ignored. This causes the evaluation of (5) to get increasingly fast as the recognition engine converges toward a single winning hypothesis.

5. Experiments

Experiments were carried out on a database of 31 synthetic 3D models of aircraft [1](see figure 1 for sample rendered images). This database was chosen in order to challenge the recognition process with a set of similar objects. During a training phase, each model was rendered from 1380 randomly selected points of view about a sphere. A small subset of the resulting images was used to construct a compact 3-dimensional feature space using principal component analysis [8]. Feature vectors were then computed for each of the training images by projection onto the resulting eigenspace. The set of possible object poses was then discretized and reduced to 46 roughly equally spaced canonical poses $$\theta_j, j \in \{1, \ldots, 46\}$$ about the viewsphere. Keeping the virtual camera at a fixed distance from the observed objects during experimentation makes the viewing position and object poses variables defined over the same domain (i.e. the sphere $$S^2$$). This establishes obvious equivalencies between different object pose and viewing position combinations; in particular,

$$p(\mathbf{d} | o_i, \theta_j, \mathbf{v}) = p(\mathbf{d} | o_i, \theta_j \oplus \mathbf{v}, \mathbf{0}), \quad (7)$$

where $$\mathbf{0}$$ denotes the origin of the global reference frame and $$\mathbf{a} \oplus \mathbf{b}$$ denotes the result of moving point $$\mathbf{a}$$ by a geometric transformation equivalent to that required to move the origin of the reference frame to point $$\mathbf{b}$$.

An appearance model is then adequately described by only the canonical likelihood distributions $$p(\mathbf{d} | o_i, \theta_j, \mathbf{0})$$, denoted by $$p(\mathbf{d} | o_i, \theta_j)$$ for clarity in the remainder of this paper. Following an approach often used in previous work [2, 3, 9, 10], the likelihood distributions were assumed to be Gaussian; that is,

$$p(\mathbf{d} | o_i, \theta_j) = \frac{1}{\sqrt{(2\pi)^{n} | \mathbf{C}_d(i, j)|}} e^{-\frac{1}{2}(\mathbf{d} - \mathbf{\bar{d}}(i, j))^T \mathbf{C}_d(i, j)^{-1}(\mathbf{d} - \mathbf{\bar{d}}(i, j))}, \quad (8)$$

where $$n = 3$$ is the dimension of the feature space and the mean $$\mathbf{\bar{d}}(i, j)$$ and covariance parameters $$\mathbf{C}_d(i, j)$$ were estimated from the sample feature vectors obtained for object $$o_i$$ for viewing positions within 15 degrees of $$\theta_j$$.

To assess the performance of the gaze planning strategy proposed in section 4, a set of experiments were performed that measured the number of views required for object recognition and pose estimation given a particular strategy. Initially, the virtual camera was positioned at the origin of the global reference frame, facing an object whose pose was selected at random. The observation thus generated was then used by the recognition engine to provide a first assessment about the class and pose of the object. All subsequent viewing positions were then chosen by the gaze planner, until the entropy of the joint posterior distribution over object class and pose fell below a confidence threshold of 0.1. Fifty such trials were performed for each object in the database where the suggested active viewpoint selection strategy was compared to a random navigation strategy.
As illustrated in figure 2, the proposed active strategy consistently recognized objects and estimated their pose using a significantly smaller number of views than random navigation, while yielding comparable accuracy. The random and active strategies yielded recognition rates of 83% and 86%, respectively, with average pose estimation errors of 1.74 degrees and 1.19 degrees.

Similar experiments were performed using (3) as a view selection criterion. (3) was evaluated using a Monte-Carlo sampling approach based on 1000 samples from the current posterior distribution $P(o_i, \theta_j|d_1, v_1, \ldots, d_i, v_i)$. Because of the complexity of this computation, only twenty trials were performed for each object. The resulting 620 trials took nearly six days to complete, while the 1550 trials performed with the proposed approach only took approximately two days. On average, the system based on (3) was able to recognize objects with 3.03 views, whereas the one based on (5) yielded an average of 2.89 views. Clearly, the suggested criterion yields comparable performance to the information theoretic measures used in previous literature and is much less expensive.

6. Conclusions

This paper presented an efficient new criterion for viewpoint selection in the context of active recognition. Empirical results on a difficult database have shown that the suggested criterion yields good recognition and pose estimation results in a short number of steps and is computationally efficient in the context of a large problem where traditional information theoretic methods are rendered intractable. The suggested approach performs better than a random navigation strategy in terms of the number of views required for object recognition and pose estimation, obtaining comparable results as with the expected mutual information criterion. 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 [10]. The suggested strategy would then provide a means of drastically reducing the cost of the training phase.

References