Robotics and Autonomous Systems 84 (2016) 31–47

Contents lists available at ScienceDirect

Robotics and Autonomous Systems

journal homepage: www.elsevier.com/locate/robot

Active multi-view object recognition: A unifying view on online feature selection and view planning

Christian Potthast *, Andreas Breitenmoser, Fei Sha, Gaurav S. Sukhatme

University of Southern California, Department of Computer Science, Los Angeles, CA 90089, USA

HIGHLIGHTS

- A unified approach to feature and viewpoint selection for multi-view object recognition is proposed.
- Online feature selection reduces the dimensionality and with that the computation time.
- View planning offers performance advantages whenever multiple views are required due to ambiguous situations or occlusions.
- Increased recognition accuracy and reduced computation cost are realized by an information-theoretic action selection framework.

ARTICLE INFO

Article history:
Received 26 May 2015
Received in revised form 14 January 2016
Accepted 27 June 2016
Available online 4 July 2016

Keywords:
Viewpoint selection
Feature selection
Multi-view object recognition
Information-theory
Active perception
Object change detection

ABSTRACT

Many robots are limited in their operating capabilities, both computational and energy-wise. A strong desire exists to keep computation cost and energy consumption to a minimum when executing tasks like object recognition with a mobile robot. Adaptive action selection is a paradigm, offering great flexibility in trading off the cost of acquiring information against making robust and reliable inference under uncertainty. In this paper, we study active multi-view object recognition and describe an information-theoretic framework that combines and unifies two common techniques: online feature selection for reducing computational costs and view planning for resolving ambiguities and occlusions. Our algorithm adaptively chooses between the two strategies of either selecting only the features that are most informative to the recognition, or moving to a new viewpoint that optimally reduces the expected uncertainty on the identity of the object. This two step process allows us to keep overall computation cost minimal but simultaneously increase recognition accuracy. Extensive empirical studies on a large RGB-D dataset, and with two different feature sets, have validated the effectiveness of the proposed framework. Our experiments show that dynamic feature selection alone reduces the computation time at runtime 2.5–6 times and, when combining it with viewpoint selection, we significantly increase the recognition accuracy on average by 8%–18% absolute compared to systems that do not use these two strategies. By establishing a link between active object recognition and change detection, we were further able to use our framework for the follow-up task of actively detecting object change. Furthermore, we have successfully demonstrated the framework's applicability to a low-powered quadcopter platform with limited operating time.

1. Introduction

Despite decades of research, object recognition from visual percepts is still an unsolved and challenging problem. In robotics especially, where many tasks build on, or directly involve, the recognition of objects, the need has grown for creating reliable and fast classification systems. Conventional approaches for object recognition are tour de force, exclusively relying in large part on complex statistical models for classification and heavy engineering of computationally-intensive features [1,2]. All too often, systems are assumed to be unlimited and their processing capabilities are neglected. This stands in stark contrast to many of today's mobile robots, which often have limited power and need to be economical with resources for computing and reasoning—being able to move and execute given tasks is their first priority!

Active object recognition [3,4] is an appealing paradigm to overcome the challenges in recognition of objects. Here, a mobile robot configures and positions its sensors to inspect a target object

* Corresponding author.
E-mail addresses: potthast@usc.edu (C. Potthast), andreas.breitenmoser@usc.edu (A. Breitenmoser), feisha@usc.edu (F. Sha), gaurav@usc.edu (G.S. Sukhatme).

http://dx.doi.org/10.1016/j.robot.2016.06.013

and recognize it as a previously learnt object. The advantage of active object recognition lies in the ability of reacting to unforeseen objects or scenes by exploratory actions. This allows to seek actively for the information that is the most useful in performing inference under uncertainty.

Adaptive action selection offers the flexibility that is required to potentially re-use the statistical modeling power of conventional approaches even under the severe (computing) constraints present on mobile robots. The general idea is to control the sensors such that only data with high information content is collected. By running inference recursively based on previously acquired information (which represents the robot's current knowledge, i.e., its belief state about the identity of the object), the collection of data can be *adapted dynamically*. Therefore, computation time can potentially be reduced, because we only process useful information and have the ability to improve our belief with each additional action.

In previous work, there are two common actions that have been explored in order to increase the recognition accuracy: first, feature selection [5], which computes a subset of informative features; second, viewpoint selection or view planning [6,3], which controls where and how many additional observations to take. However, previous approaches have not looked at the potential of reducing computation time due to feature selection. Moreover, existing work does not exploit the full potential of combining these two strategies and applying feature and viewpoint selection in concert.

The main idea of our work hinges on the following intuition: *Is it possible to reduce overall costs for recognition by relying on feature subsets gathered from multiple views*, in lieu of the full feature set, potentially taken from a single view, using a complex statistical model for classification? More concretely, in this paper, we present an information-theoretic multi-view object recognition framework that unifies the two strategies of *feature selection* and *viewpoint selection*. We aim at exploiting local information by selecting additional features at the current viewpoint and further exploring the object by selecting new viewpoints around it. The described framework optimizes the viewpoint selection by reducing ambiguities and selecting features dynamically, such that only useful and informative features are extracted and incorporated in the recognition process. Unlike standard feature selection, the two strategies are symbiotic in our algorithm. Feature selection is an iterative process, thus at any given time, the selected feature depends on the trajectory of the past viewpoints while simultaneously affecting the future viewpoint(s).

We perform extensive evaluations on a large RGB-D camera dataset, which is composed of single object scenes. To evaluate our framework, we conduct experiments with two different feature sets, one set consisting of *simple features* designed by us, and one set of more complex *kernel descriptors* proposed by [1]. The use of the simple features shows that, in combination with feature selection and view planning, our framework can compete with state-of-the-art single-view classification systems. In contrast, using the kernel descriptors shows that our framework is capable of dealing with more advanced features as well.

We further show the advantages of combining intelligent view planning and online feature selection. Online feature selection reduces the computation time fivefold for the simple features and 2.5-fold for the kernel features at runtime. Furthermore, in combination with view planning, we are able to increase the recognition accuracy significantly by resolving object ambiguities. In the case of the simple features, we achieve an increase in accuracies by 8%–15%, while in case of the kernel descriptors the increase is about 8%. Overall, our results show pose estimates within $\pm 45^{\circ}$ of the ground truth and object recognition accuracies of over 90% when using the simple features in rather challenging object recognition experiments. Using the kernel descriptors, we improve the object recognition accuracy to over 98%. As proof-of-concept, we eventually implemented and tested the proposed approach on a custom-built quadcopter robot. The quadcopter is equipped with an RGB-D camera and collects multiple views by flying around a target object, which allows to recognize the object.

Additionally, we show how our active multi-view recognition system can be used for *object change detection*. Given a prior map of the environment, the task is to determine whether objects in the environment have changed. This can either mean that the object has changed completely (i.e., it belongs to a different object class), or the object has changed its orientation only (i.e., it belongs to the same class but its rotation differs in yaw). Our results show that on average our framework can detect such changes in the environment with an accuracy of larger than 90%.

The remainder of the article is organized as follows. Section 2 presents related work on active object recognition and pose estimation, including relevant approaches in feature selection and viewpoint planning. The problem formulation and proposed solution of our active multi-view object recognition approach is detailed in Section 3. Section 4 introduces the unified selection of actions, which is applied to both feature selection in Section 5 and viewpoint selection in Section 6. The object recognition experiments and results are discussed in Section 7, and a conclusion is provided in Section 8.

2. Related work

Prior works on active perception go back as far as the seminal paper [7]. "Active" can either refer to adjustments of the sensor parameters themselves, e.g., adaptively changing the focal length of a camera, as in [7,4], or mean that the entire sensor moves together with the underlying platform. The latter is particularly common in robotics, where robots plan new observation locations and viewing directions actively [8,9,3]. This is known as robotic view planning. [6] presents a comparison of several representative approaches to view planning from literature. View planning can be formulated as a purely geometric problem [10,11] but as well relates to robotic exploration, where information-theoretic strategies have become state-of-the-art solutions [12,13]. The selection of optimal viewpoints for a robot can be useful in various applications, such as for the inspection of underwater structures [14], for object search in indoor environments [15,16] or detecting objects on a table top [3]. Moreover, methods for view planning and active object recognition provide the basis for operations like object sorting [17] and manipulation [13,18], acquisition of object representations [19] or change detection (we will show in the end of Section 7 how our object recognition framework naturally extends to perform object change detection).

The common goal of object recognition approaches is to recognize objects and estimate their poses [4,8,15,20,21]. [4] presents an active object recognition framework that, similar to ours, relies on a sequential decision process and mutual information as information-theoretic measure for the selection of actions. [8] selects multiple viewpoints by using the Jeffrey's divergence instead of maximizing the mutual information. [15] combines active object recognition with Simultaneous Localization and Mapping (SLAM) techniques to estimate object positions and link semantic with spatial information. [20] extends a single-view to a multi-view object recognition method for increased efficiency by applying multi-step optimization and extracting SIFT features from camera images. [21] performs object recognition and object detection by using different classifiers and several shape and visual features.

To cope with objects that appear ambiguous in single views, [22–24,20] propose multi-view object recognition frameworks