Towards Interactive Object Recognition

Karol Hausman Chet Corcos Jörg Müller Fei Sha Gaurav S. Sukhatme

Department of Computer Science, University of Southern California, Los Angeles, CA, USA
{hausman, corcos, joerg.mueller, feisha, gaurav}@usc.edu

I. Introduction

Object recognition is a key component of service robots for finding and handling objects. Current state-of-the-art object recognition systems recognize objects based on static images [7, 8]. However, these systems prove limited in cases when objects are in ambiguous orientations or distinctive features are hidden, e.g., due to the pose of the object.

A popular approach to tackle this problem is active perception [1, 3], where the robot intelligently moves its camera to reveal more information about the scene. However, there are cases where this approach will fail because distinctive features are hidden, for example, on the bottom side of the object (see Fig. 1). These cases are particularly common in cluttered environments, where features might be occluded not only due to the pose of the object but also by other items in the scene. It has been recently studied in the area of interactive perception that interacting with the scene exposes new possibilities to tackle common perception problems. This paper addresses both challenges—selecting an object of a cluttered scene for manipulation and picking the optimal movement of this object—in an information-theoretic way to improve interactive perception methods.

Interacting with a scene to improve perception by revealing informative surfaces has been particularly explored in the area of segmentation. Examples are: interactive segmentation of rigid objects being moved by a robot [5], segmentation of articulated objects [4], and disambiguation of segmentation hypothesis [2]. However, none of these approaches reason about what actions to take in order to achieve the goal.

In this work we introduce a probabilistic method for choosing object manipulation actions to optimally reveal information about objects in a scene based on robot's observations. To the best of our knowledge, the problem of interactive object recognition has not been addressed before. Our approach determines the optimal action for a robot to interact with objects and adjust their pose to reveal discriminative features for determining their identity. In the ambiguous book example (see Fig. 1), this means flipping the book over and observing the cover, which results in more confident recognition. Our method is based on a probabilistic graphical model for feature-based object and pose recognition. By inferring posterior distributions of object probabilities conditioned on all previous actions and observations, our approach enables a robot to select the optimal action to reduce the uncertainty of the object.

The key contributions of this approach are: (a) it presents a probabilistic action selection model that reasons about the most informative action and (b) it uses a probabilistic object recognition model that is indifferent of the feature type.

Fig. 1. Top-left: The service robot PR2 trying to recognize a book based on its back. The database of objects consists of book 1 (top-right, NE and NW) and book 2, (top-right, SE and SW) that look the same from the back. PR2 takes the optimal action in order to recognize which book it is. In this case it means it flips it over (bottom-left, bottom-right).

II. Approach

Our approach chooses actions that minimize the uncertainty about an object being observed. We introduce a feature-based observation model that is used for probabilistic object recognition. We extend this model into a temporal graphical model to incorporate actions. Finally, we propose an expected entropy measure to find the optimal action that will minimize the uncertainty of the object.

A. Probabilistic Graphical Model

1) Observation Model: We use an observation model $p(\mathbf{F}|o,p)$ where object and pose result in the appearance of specific features that are observed by the robot. This graphical model is shown in dotted lines in Fig. 2. The model consists of N discrete objects, $O \in \{o_1, o_2, ..., o_N\}$ in I discrete poses $P \in \{p_1, p_2, ..., p_I\}$. We model M features $\mathbf{F} = \{f_1, ..., f_M\}$ where $\mathbf{F}$ is a set of continuous random variables f_i. This model

assumes features are conditionally independent given an object and its pose.

2) Object Recognition: The posterior of the object-pose is given by Eq. (1) with some prior $\mathrm{p}(o, p)$ and the observation model $\mathrm{p}(\mathbf{F}|o, p)$.

$$\mathrm{p}(o, p|\mathbf{F}) = \frac{\mathrm{p}(o, p) \cdot \mathrm{p}(\mathbf{F}|o, p)}{\sum_{n,i} \mathrm{p}(\mathbf{F}|o_n, p_i) \cdot \mathrm{p}(o_n, p_i)} \tag{1}$$

3) Interactive Object Recognition: To model actions, the object-recognition subgraph is extended into a temporal graphical model. For each pose, actions are modeled as I relative pose transformations including the *stay* action. In this model, the next pose P_{t+1} is dependent only on the previous pose P_t and the previous action A_t. This results in the graphical model shown in Fig. 2.

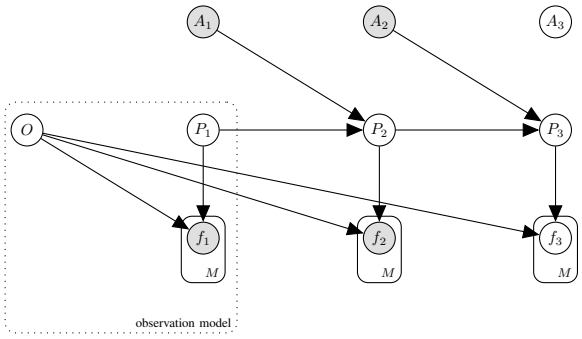

Fig. 2. Probabilistic graphical model for interactive object recognition.

The posterior at time $t+1$ given the entire history of observations and actions is a recursive Bayesian update of the posterior at time t given in Eq. (2).

$$\mathrm{p}(o, P_{t+1}|\mathbf{F}_{1:t+1}, A_{1:t}) = \frac{\sum_{P_t} \mathrm{p}(o, P_t|\mathbf{F}_{1:t}, A_{1:t-1})\mathrm{p}(\mathbf{F}_{t+1}|o, P_{t+1})\mathrm{p}(P_{t+1}|P_t, A_t)}{\sum_{P_t, P_{t+1}, O} \mathrm{p}(O, P_t|\mathbf{F}_{1:t}, A_{1:t-1})\mathrm{p}(\mathbf{F}_{t+1}|O, P_{t+1})\mathrm{p}(P_{t+1}|P_t, A_t)} \tag{2}$$

4) Optimal Action Selection: We define the optimal action for object recognition as moving an object into a pose in which the next observation minimizes the uncertainty of the object. This results in a minimum entropy of the distribution of posterior object prediction probabilities.

Because we haven't observed $\mathbf{F}_{t+1}$, we must compute the *expected* entropy of the posterior in Eq. (2). The optimal action is selected as the action which minimizes the expected entropy of object prediction posteriors across all potential actions:

$$A_t^* = \underset{A_t}{\operatorname{argmin}}\, \mathbb{E}_{\mathbf{F}_{t+1} \sim \mathrm{p}(\mathbf{F_{t+1}}|\mathbf{F}_{1:t}, A_{1:t})} \mathrm{H}\left[\, O|\mathbf{F}_{1:t+1}, A_{1:t}\,\right] \tag{3}$$

B. Implementation

1) Observation Model: Each feature in the model has an associated type j and a value or descriptor with which to compute a similarity or matching error $\mathcal{E}^j(\cdot, \cdot)$ with respect to another feature of the same type. Object and pose are predicted using a model $\mathrm{p}(f|o, p)$ derived from matching errors between observed feature values, $\mathbf{F}_{obs}$ and the set of reference feature values of the model, $\mathbf{F}$. The features of the model are selected as the set of all unique features from all objects and poses observed in an ideal setting. Given an observation, $\mathbf{F}_{\text{obs}}$, the best matching error e with respect to a feature in the model $f^j \in \mathbf{F}$ is given by Eq. (4).

$$e(f^j) = \min_{f^j_{\text{obs}} \in \mathbf{F}_{\text{obs}}} \mathcal{E}^j(f^j, f^j_{\text{obs}}) \tag{4}$$

For our model, we used SIFT [6] features and approximate the distribution of $e(f^j)$ by a normal distribution.

2) Optimal Action Selection: To efficiently compute the expected entropy given in Eq. (3), the posterior distribution is sampled for each action. First, the evidence given in Eq. (5) is sampled.

$$\mathrm{p}(\mathbf{F}_{t+1}|\mathbf{F}_{1:t}, A_{1:t}) = \sum_{P_t, P_{t+1}, O} \mathrm{p}(\mathbf{F}_{t+1}|O, P_{t+1})\mathrm{p}(P_{t+1}|P_t, A_t)\mathrm{p}(O, P_t|\mathbf{F}_{1:t}, A_{1:t-1}) \tag{5}$$

This distribution can be sampled trivially by first sampling object-poses based on the discrete distribution defined by the previous posterior, $\mathrm{p}(O, P_t|\mathbf{F}_{1:t}, A_{1:t-1})$. Then, for each sampled object-pose, a sample representing a potential next observation is drawn from the feature likelihood distribution $\mathrm{p}(\mathbf{F}_{t+1}|O, P_{t+1})$. In our experiment, we assume a perfect actuator, i.e. $\mathrm{p}(P_{t+1}|P_t, A_t) \in \{0, 1\}$. Thus, given an action and a pose, the next pose can be computed deterministically.

The next posterior is computed for each sample by Eq. (2). The posterior object probability is computed by marginalization given in Eq. (6).

$$\mathrm{p}(O|\mathbf{F}_{1:t+1}, A_{1:t}) = \sum_{P_{t+1}} \mathrm{p}(O, P_{t+1}|\mathbf{F}_{1:t+1}, A_{1:t}) \tag{6}$$

The entropy of the posterior object probabilities is computed for each sample and then averaged to give the expected entropy of the object posterior. The expected entropy is computed for each potential action and the optimal action is selected according to Eq. (3).

III. Experimental Results

We evaluated the proposed approach on a dataset consisting of $N = 4$ books in $I = 4$ poses. We used two pairs of books which are ambiguous on the back and unambiguous on the front. Fig. 3 shows the covers of all the books used for the experiment. All poses are presented in Fig. 4.

$M = 654$ unique features were extracted from a set of ideal images of each object-pose pair. We recorded 100 training samples for each object-pose pair to learn the likelihood distribution $\mathrm{p}(f|o, p)$. For the ambiguous cases, we used the same training images.

Our experimental setup consists of an RGB camera and one of the books. In our preliminary experiment, all the actions were executed by a human.

Fig. 3. Books from the cover side used for the experiment. Two first books and two last books look the same from the back side.

Fig. 4. All the poses used for object-pose recognition. Please note the visibility of the spine.

A. Object Recognition

In order to evaluate the object recognition model, we trained the model on 80 samples and held out 20 samples for cross validation. The average prediction accuracy for the unambiguous cases is 99.67% for the training data and 93.75% for the cross validation data. We did not include the ambiguous poses in the cross validation results because these ambiguous cases were designed to cause static object recognition to fail.

B. Action Selection

An action selection experiment is represented by the decision tree in Fig. 5. The ambiguous back of the book pose was observed as shown in Fig. 6 (top-left). As expected, the posterior probabilities were split between the two ambiguous books shown in Fig. 6 (top-right).

Of the four actions, staying and rotating result in similarly ambiguous poses resulting in an expected entropy of 0.7. Flipping the book over, with and without rotating, lead to similarly unambiguous poses with an expected entropy of 0. After flipping the book, the robot observes the cover

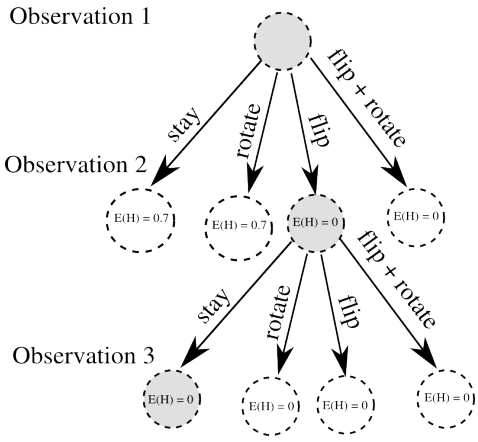

Fig. 5. Decision tree based on the action selection algorithm. Each node in the tree represents the expected entropy of the posterior probability for a given action. Colored nodes indicate the choice of the action that results in the minimum expected entropy.

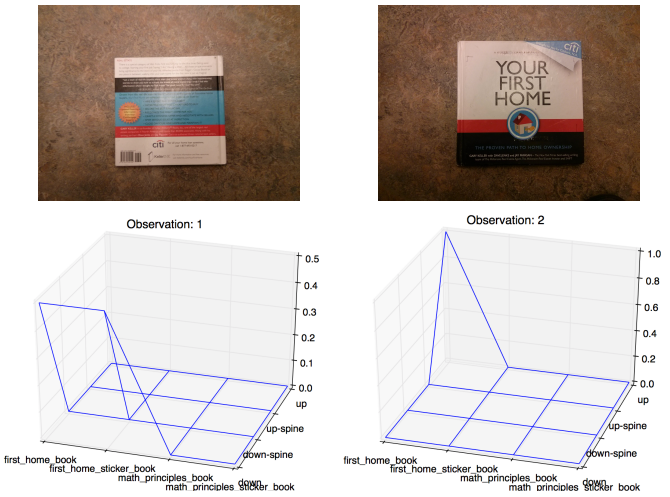

Fig. 6. Left: observed image for the first observation (top) and the corresponding posterior probability of object and pose after the first observation (bottom). Right: Analogous graphs after the second observation (i.e. after the action was taken).

(Fig. 6 bottom-left) and predicted the correct object with 100% certainty (Fig. 6 bottom-right).

IV. Conclusions

We have presented a probabilistic framework for interactive object recognition. We formulated a minimum expected entropy principle for determining the optimal action to reduce uncertainty in object recognition. A preliminary experiment on the ambiguous book problem shows encouraging results.

There are several areas for future work in this domain. We believe that loosening our constraints on discrete poses with perfect actions into continuous poses and noisy actions will enable this work to be very useful in cluttered environments.

References

[1] N. Atanasov, B. Sankaran, J. Le Ny, T. Koletschka, G.J. Pappas, and K. Daniilidis. Hypothesis testing framework for active object detection. In *IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2013.

[2] N. Bergström, C.H. Ek, M. Björkman, and D. Kragic. Scene understanding through interactive perception. In *8th Int. Conf. on Computer Vision Systems (ICVS)*, 2011.

[3] Geoffrey A Hollinger, Urbashi Mitra, and Gaurav S Sukhatme. Active classification: Theory and application to underwater inspection. *arXiv preprint arXiv:1106.5829*, 2011.

[4] D. Katz and O. Brock. Interactive segmentation of articulated objects in 3d. In *Workshop on Mobile Manipulation at ICRA*, 2011.

[5] J. Kenney, T. Buckley, and O. Brock. Interactive segmentation for manipulation in unstructured environments. In *IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2009.

[6] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

[7] J. Tang, S. Miller, A. Singh, and P. Abbeel. A textured object recognition pipeline for color and depth image data. In *IEEE Int. Conf. on Robotics & Automation (ICRA)*, 2012.

[8] J. van de Weijer and F.S. Khan. Fusing color and shape for bag-of-words based object recognition. In *Computational Color Imaging*. 2013.