Learning Object and State Models for AR Task Guidance

William Hoff*, Hao Zhang*
Division of Computer Science, EECS Department, Colorado School of Mines, Golden, CO 80401

ABSTRACT

We present a method for automatically learning object and state models, which can be used for recognition in an augmented reality task guidance system. We assume that the task involves objects whose appearance is fairly consistent, but the background may vary. The novelty of our approach is that the system can be automatically constructed from examples of experts performing the task. As a result, the system can be easily adapted to new tasks. The approach makes use of the fact that the key features of the object are consistently present in multiple viewing instances; whereas features from the background or irrelevant objects are not consistently present. Using information theory, we automatically identify the features that can best discriminate between object states. In evaluations, our prototype successfully recognized object states in all trials.

Keywords: State recognition, egocentric vision, augmented reality, task guidance.

Index Terms: I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Object Recognition

1 INTRODUCTION

Augmented reality (AR) has the potential to improve the effectiveness of personnel in performing tasks such as maintenance, repair, and the operation of complex equipment. Research has demonstrated that using AR systems for maintenance and assembly tasks results in improved performance, in terms of reduced task time [3] and reduced number of errors [5]. One major issue preventing the widespread adoption of AR task guidance systems is the expense of developing them. Although the hardware is relatively inexpensive, each application requires a large amount of manual effort to develop. For example, in some approaches markers are attached to the objects and their positions are manually determined [6].

Recently, methods based on egocentric vision have been developed to automatically learn the appearance of objects. For example, [1] discovers task relevant objects by clustering features extracted from images taken by multiple people interacting with the objects. In addition to recognizing an object, we also want to recognize the "state" of the object. By "state", we mean a configuration of the object that is determined by the presence or position of subparts. For example, a printer/fax machine may have its front cover closed or open, as illustrated in Figure 1. Sometimes the visual differences between states can be subtle, such as the presence or absence of a screw, or the position of a switch.

In this paper, we outline a novel approach for developing AR task guidance systems for a wide variety of tasks, with little or no programming effort. Due to space limitations, we focus here on the problem of learning to recognize object and state, although our system also includes learning a model of the task workflow, and the user interface. We use a novel method based on information theory in order to find the most discriminative features. The discriminative features allow the system to distinguish between object states whose appearance can be very similar.

*e-mail: {whoff, hzhang}@mines.edu

Figure 1: Two states of a printer, as viewed by a person wearing a head mounted camera.

2 OVERALL APPROACH

Our approach uses the metaphor of an expert training a novice. A human expert, wearing the AR system, first demonstrates the task as if instructing a human novice. Similar to [2], an expert verbally describes the objects he or she is working with and the actions being performed. For example, the expert might say "This is the drum unit, containing the toner cartridge", which is a description of the state of the objects in the scene. The expert might say "I am now pulling out the toner cartridge from the drum unit," which is the action they are performing. The result is a video stream that is annotated with names of objects in a particular state.

We capture data from a small number (on the order of 5 to 10) of experts performing the task. To automatically learn object appearance, we use the fact that we have videos of the same object taken in different scenes. The appearance of the object is fairly consistent, but the background changes between scenes. For example, Figure 2 shows images from two different videos of the printer maintenance task. By finding the common features between the images, we can identify the features that belong to the printer, and ignore features from the background.

Figure 2: The same object in different scenes.

3 DETAILED DESCRIPTION

First of all, we process the video segments and automatically extract keyframes [4], which are a set of images taken from cameras that are sufficiently far apart in pose. The idea is to avoid saving and processing redundant images, where there is little or no change in appearance. The keyframes are put into a database, along with a label for the object state. At this point, we still have not segmented the object, or identified any characteristics that distinguish one state from another.

Keypoints (e.g., SIFT) from each training image are matched to every other training image. We verify candidate feature correspondences by fitting a fundamental matrix (using RANSAC to find inliers) to the points. Figure 3 shows the most salient features for two

of the database images; meaning those features that had the highest probability of finding a match to a feature in another database image. As can be seen, features on the object are clearly identified, with no features in the background. Thus, with no hand segmentation we have automatically learned features belonging to the object.

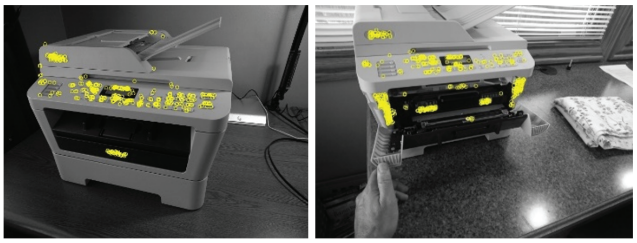

Figure 3: Salient features on the object, automatically identified.

We also calculate the conditional probability of each feature to have a match to an image of a specific state. Namely, we compute the conditional probability $p(m_j = 1|s_i = k)$ and $p(m_j = 0|s_i = k)$, where s_i is the object state in the i-th image, and m_j is the outcome of matching feature j to the i-th image; i.e., $m_j = 1$ if a match is found to another feature, and $m_j = 0$ if not. Next, we determine the ability of each feature to discriminate between states. The best features to use are those that most reduce the entropy of the set of images. Specifically, we calculate the entropy of the original set of database images as

$$H_{parent} = \sum_k p_k \log_2(p_k), \tag{1}$$

where p_k is the probability of an image being in state k. Next, each feature conceptually partitions the images into two child sets. If a feature has a match in another database image, then the other image goes into one child set (A), and if not, the image goes into the other child set (B). We calculate the entropy of each child set, and form the weighted average of the child sets:

$$H_{children} = p(m_j = 1)H_A + p(m_j = 0)H_B, \tag{2}$$

where m_j is the outcome of matching feature j; i.e., $m_j = 1$ if a match is found for this feature, and $m_j = 0$ if not.

Figure 4: Discriminative features, as computed by information gain. The presence or absence of these features is helpful for determining which state the object is in.

The "information gain" for a feature is the difference between the entropies of the parent set and the children. Features with high information gain are the most useful features for discriminating between states. Figure 4 shows the features with the highest information gain, for the two database images used as examples.

As a second example, the proposed system was applied to the task of changing a car air filter. Figure 5 demonstrates two object states from this task. In one image, the air filter is present; in the other it has been removed. As can be seen the system has identified discriminative features on the filter.

Once training is done, the system can recognize the object in a query image and determine which state it is in. To do this, features are extracted from the query image and matched to the images in the database. The best matching database images (as determined by the number of feature matches) form a set of candidate images.

Each candidate image computes the likelihood of the query image to be in a particular state, using Bayes' rule:

$$p(s_q = k|\{m_j\}) = \frac{p(\{m_j\}|s_q = k)p(s_q = k)}{p(\{m_j\})}, \tag{3}$$

where $\{m_j\}$ is the set of matches of the points in the candidate image to the query image, and $p(s_q = k)$ is the *a priori* probability for the query image to be in state k. We utilize only discriminative features for this calculation (in our prototype we use features that have information gain $\geq I_{max}/2$).

Finally, the likelihoods from each candidate image are combined by multiplying (in our implementation we add their logs) to obtain a final probability for the query image to be in each state. The candidate image with the highest probability is taken to be the correct match for object and state. In evaluations on the printer task and the air filter task, our prototype successfully recognized the objects and states in all trials.

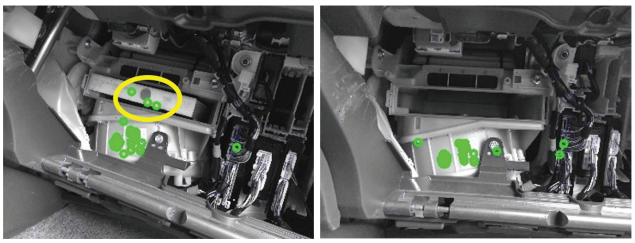

Figure 5: The three discriminative features on the air filter (circled in yellow on the left image) can help distinguish this state from the case where the air filter is missing (right).

4 Conclusion

Our approach can learn to recognize objects (and object states) using no user guidance, other than capturing video from a small number of experts performing the task. No reprogramming is necessary to apply the system to a new maintenance task. The experts need only announce the individual steps in the task as they go, as if they were instructing a novice. We use a novel method based on information theory in order to find the most discriminative features and identify object states.

Acknowledgements

This work was partially supported by a gift from DAQRI, LLC.

References

[1] D. Damen, T. Leelasawassuk, and W. Mayol-Cuevas. You-do, i-learn: Egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. *CVIU*, 149:98–112, 2016.

[2] A. Fathi. Learning descriptive models of objects and activities from egocentric video. PhD thesis, Georgia Institute of Technology, 2013.

[3] S. Henderson and S. Feiner. Exploring the benefits of augmented reality documentation for maintenance and repair. *TVCG*, 17(10):1355–1368, 2011.

[4] G. Klein and D. Murray. Parallel tracking and mapping for small ar workspaces. In *ISMAR*, 2007.

[5] S. Webel, U. Bockholt, T. Engelke, N. Gavish, M. Olbrich, and C. Preusche. An augmented reality training platform for assembly and maintenance skills. *RAS*, 61(4):398–403, 2013.

[6] J. Zauner, M. Haller, A. Brandl, and W. Hartmann. Authoring of a mixed reality assembly instructor for hierarchical structures. In *ISMAR*, 2003.