Attribute Based Object Recognition by Human Language (Extended Abstract)

Zhe Zhao
University of Science and Technology of China
zhaozhe@mail.ustc.edu.cn

Jiongkun Xie
University of Science and Technology of China
devilxjk@mail.ustc.edu.cn

Xiaoping Chen
University of Science and Technology of China
xpchen@ustc.edu.cn

ABSTRACT

Over the last years, the robotics community has made substantial progress in detection and 3D pose estimation of known and unknown objects. However, the question of how to identify objects based on language descriptions has not been investigated in detail. While the computer vision community recently started to investigate the use of attributes for object recognition, these approaches do not consider the task setting typically observed in robotics, where a combination of colors, shapes, materials might be used in referral language to identify specific objects in a scene. In this paper, we introduce an approach for identifying objects based on natural language containing the attributes of the object. Our experiments show that by using the attributes mentioned in the referral language it is indeed possible to build a learning object detection system that does not require any training images of the target classes.

1. INTRODUCTION

Identifying objects in complex scenes is a crucial capability for an autonomous robot to understand and interact with the physical world and be of use in everyday life scenarios. Over the last years, the robot community has made substantial progress in object detection and pose estimation. All work on object recognition assumes that each object has a unique name. However, this is not the way how humans identify objects. People often use the object's attributes to describe the object. For instance, a person might say "I want the red apple" or "Please give me some food". The first situation requires a color attribute, and the material attribute is needed in the second situation.

In this paper, we introduce an approach for identifying objects based on object's attributes. A robot is given some objects and a sentence which contain the attribute of the target object. The robot has to identify the target object based on the attributes described by the sentence. We tackle the problem by introducing an attribute-based classification. It performs object detection based on a human-specified high-level description of the target objects instead of training images. The description consists of arbitrary semantic attributes, like material, color or even shape information. Because such properties transcend the specific learning task at hand, they can be pre-learned, e.g.from image datasets unrelated to the current task. Afterwards, new classes can be detected based on their attribute representation, without the need for a new training phase. Also, we can get the attribute of the target object by analyzing the sentences human gives.

To evaluate our approach, we use the RGB-D dataset[1] developed by Lai and colleagues. We also create the test dataset which contains multiple objects and simulates a task in which a person can command a robot to pick up one object from the objects. Experiments demonstrate that our system can identify objects well by human command.

Appears in: *Alessio Lomuscio, Paul Scerri, Ana Bazzan, and Michael Huhns (eds.), Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2014), May 5-9, 2014, Paris, France.*

2. VISUAL ATTRIBUTE LABELING

An object could be described in various ways. They are usuall described in natural language, providing some visual attributes of the object. For example, a description of a plate would be: "*The plate, which is ellipse, is made of metal*". Extracting some visual attributes from the natual language sentences could help the visual recognition. The goal of visual attribute extraction is to find out which parts of the object description providing the information of attributes and what kind of attributes they are respectively. Our approach to extracting visual attributes first chunks together the words in an object description to get its shallow syntactic structure. Then a tagger takes as input the chunks and produces a sequence of visual attribute tags. By these tags, the visual attributes are extracted.

Following [2], we adopt the IOB2 representation for the visual attributes. In our case, each visual attribute classifier corresponds to two IOB2 tags which are prefixed with 'B_' and 'I_' denoting respectively the begining and inside of an fragment. For example, 'B_ellipse' starts a new segmentation in the description for the visual attribute *ellipse*. The next words tagged with 'I_ellipse' are inside the segmentation. The segmentation continues until the word is not tagged with 'I_ellipse'. In addition, there is an extra tag 'O' representing words that are not corresponding to any visual attributes. Fig. 1 shows the example of our IOB2 representation.

The/O plate/O ,/O which/O is/O ellipse/B_ellipse ,/O is/O made/O of/O metal/B_metal ./O

Figure 1: An example of IOB2 representation for visual attribute

[1] `http://rgbd-dataset.cs.washington.edu/`

Our approach to visual attribute tagging takes as input a sequence of words with part of speech and syntactic chunks then determine the most possible sequence of visual attribute tags. We make use of the *conditional random field* to model such sequence prediction:

$$\mathbf{y}^* = F(\mathbf{x}) = \arg\max_{\mathbf{y}\in\mathbf{GEN}(\mathbf{x})} \frac{\exp[\theta \cdot f(\mathbf{x}, \mathbf{y})]}{\sum_{\mathbf{y}'\in\mathcal{Y}} \exp[\theta \cdot f(\mathbf{x}, \mathbf{y}')]}$$

where $\mathbf{x}$ is a set of input chunks, $\mathbf{y}^*$ is the best scored result, and the function **GEN** enumerates all candidates of visual attribte tag on $\mathbf{x}$. We employ the Illinois Chunker[2] to generate the input chunks $\mathbf{x}$.

3. ATTRIBUTE-BASED CLASSIFICATION

Attribute-based classification models object classes relative to an inventory of descriptive attributes. For a given class, each attribute can be either active or inactive, resulting in a characteristic association signature for that class.

Following the probabilistic formulation of the DAP model in [1], let $a_y = (a_1^y, ..., a_m^y)$ be a vector of binary associations $a_m^y \in \{0, 1\}$ between attributes a_m and training object classes y. A classifier for attribute a_m, trained by labeling all images of all classes for which $a_m^y = 1$ as positive and the rest as negative training examples, can provide an estimate of the posterior probability $p(a_m|x)$ of that attribute being present in image x. Mutual independence yields $p(a|x) = \prod_{m=1}^{M} p(a_m|x)$ for multiple attributes.

In order to transfer attribute knowledge to an unknown class z, we again assume a binary vector a^z for which $p(a|z) = 1$ for $a = a^z$ and 0 otherwise. The posterior probability of class z being present in image x is then obtained by marginalizing over all possible attribute associations a, using Bayes' rule $p(z|a^z) = \frac{p(a^z|z)p(z)}{p(a^z)} = \frac{p(z)}{p(a^z)}$:

$$p(z|x) = \sum_{a\in\{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m|x)^{a_m^z} \quad (1)$$

Assuming identical class priors $p(z)$ and a factorial distribution for $p(a) = \prod_{m=1}^{M} p(a_m)$, we obtain

$$p(z|x) \propto \prod_{m=1}^{M} \left(\frac{p(a_m|x)}{p(a_m)}\right)^{a_m^z} \quad (2)$$

Attribute priors can be approximated by empirical means over the training classes $p(a_m) = \frac{1}{K}\sum_{k=1}^{K} a_m^{y_k}$. Classifying an image x according to test classes z_L uses MAP prediction:

$$f(x) = \max_{l=1,...,L} \prod_{m=1}^{M} \left(\frac{p(a_m|x)}{p(a_m)}\right)^{a_m^{z_l}} \quad (3)$$

4. EXPERIMENTAL EVALUATION

4.1 Attribute Labeling

To evaluate our approach to visual attribute extraction, we had collected 1000 sentences that describe an object from differnt aspects, e.g., color, shape, and material. Each description was syntactically chunked by the Illinois Chunker and manually labeled with the visual attribute tags. The L-BFGS algorithm with a Gaussian prior smoothing was employed to estimate the parameters of the model. Our experiments used 5-fold cross validation. We measured the performance of visual attribute extraction by using the *precision* (percentage of returned segmentations that were correct), *recall* (percentage of correct segmentations actually presented in the input), and *F-measure* (harmonic mean of precision and recall). The segmentation was correct if its corresponding sequence of visual attribute taggs are matched the correct ones. The experiment results of our approach to visual attribute extraction are shown in Table 1.

Table 1: Experiment results

	Precision	Recall	F_1
Test 1	100.00%	33.43%	50.11%
Test 2	100.00%	34.76%	51.58%
Test 3	100.00%	35.86%	52.79%
Test 4	100.00%	37.09%	54.11%
Test 5	100.00%	36.53%	53.51%
Average	100.00%	35.53%	52.43%

The experiment results showed our approach had a precise prediction on visual attributes. However, the recall was low indicating that our approach could not retrieve most visual attributes. There was a lot of room for the improvement. To test our system, we selected 16 categories from the RGB-D Object Dataset. First we show how well we can assign attributes and use them to describe objects. We examine the performance of using the attribute based representation in the traditional naming task and demonstrate the zero-shot ability.

4.2 Detect Known and Unknown Objects

There are two main protocols for attribute prediction: "within category" predictions, where train and test instances are drawn from the same set of classes, and "across category" predictions where train and test instances are drawn from different sets of classes. The experiments used 5-fold cross validation. The result is shown in Table 2.

Table 2: Experiment results

	known objects	unknown objects
Test 1	93%	76%
Test 2	91%	65%
Test 3	92%	80%
Test 4	88%	77%
Test 5	89%	76%
Average	90.6%	74.8%

5. REFERENCES

[1] C. H. Lampert, H. Nickisch, and S. Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *CVPR*, pages 951–958, 2009.

[2] E. F. T. K. Sang and J. Veenstra. Representing Text Chunks. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, pages 173–179, University of Bergen, Bergen, Norway, 1999.

[2] `http://cogcomp.cs.illinois.edu/page/software_view/13`