Towards Adaptive Object Recognition for Situated Human-Computer Interaction

[Position Paper]

Kate Saenko
Computer Science and Artificial Intelligence Laboratory
Massachusets Institute of Technology
Cambridge, MA, USA
saenko@csail.mit.edu

Trevo
Computer Science a
Lab
Massachusets Ins
Cambridg
trevor@c

ABSTRACT

Object recognition is an important part of human-computer interaction in situated environments, such as a home or an office. Especially useful is category-level recognition (e.g., recognizing the class of chairs, as opposed to a particular chair.) While humans can employ multimodal cues for categorizing objects during situated conversational interactions, most computer algorithms currently rely on vision-only or speech-only recognition. We are developing a method for learning about physical objects found in a situated environment based on visual and spoken input provided by the user. The algorithm operates on generic databases of labeled object images and transcribed speech data, plus unlabeled audio and images of a user refering to objects in the environment. By exploiting the generic labeled databases, the algorithm would associate probable object-referring words with probable visual representations of those objects, and use both modalities to determine the object label. The first advantage of this approach over visual-only or speech-only recognition is the ability to disambiguate object categories using complementary information sources. The second advantage is that, using the additional unlabeled data gathered during the interaction, the system can potentially improve its recognition of new category instances in the physical environment in which it is situated, as well as of new utterances spoken by the same user, compared to a system that uses only the generic labeled databases. It can achieve this by adapting its generic object classifiers and its generic speech and language models.

1. INTRODUCTION

The ability to categorize places and objects in the immediate environment is important for many HCI applications, including mobile robotic assistants for the elderly and the disabled. One approach to categorization is through *image*-based recognition, which involves training a classifier for each object category offline, using manually labeled images. However, obtaining labeled images for a large variety of objects requires a lot of manual effort. Alternatively, a robot can learn about its surroundings from interactions with the user, such as in the "home tour" scenario [4], where the user points to objects around the room and describes them verbally, e.g., "this is my pen". However, such "show-and-tell" systems do not use any prior knowledge of object category appearance, but rather use the output of the speech recognizer to determine the object. If the spoken description is misrecognized, an incorrect object label may be assigned to the input image (e.g., "pan", instead of "pen".) Such systems are also limited to visually simple objects and backgrounds, since they rely on regions of solid color to segment objects from background. Finally, they learn only from the examples provided by the user. This places a burden on the user to show the robot many different objects of the same category, if the system is to generalize to unseen instances.

We propose a new approach, which overcomes the above limitations by combining speech- and image-based category recognition. The motivation comes from the fact that humans use multiple modalities to understand which object category is being referred to, simultaneously interpreting speech and visual appearance. The approach consists of two parts: *disambiguation* and *adaptation*. It starts with generic image and speech classifiers trained on offline databases that are not tailored to the specific environment or user. Disambiguation combines the classifiers to obtain the correct object labels for new image and utterance pairs provided by the user. Even if the individual classifiers are weak, together they can better constrain recognition. For example, if the ASR component returns "paris" as the most likely word, and "purse" as the fifth, the visual classifier might be able to rule out the incorrect choices based on the corresponding image context (see Figure 1). The same intuition applies in the other direction, with speech disambiguating confusable visual categories.

Once new labels have been determined, the adaptation step uses them to adapt the classifiers on the resulting environment- and user-specific data. This approach can be thought of as a form of knowledge transfer from one problem (classifying objects in the offline database) to another (classifying objects in the situated environment). Thus, given a large labeled offline database and a small number of labeled

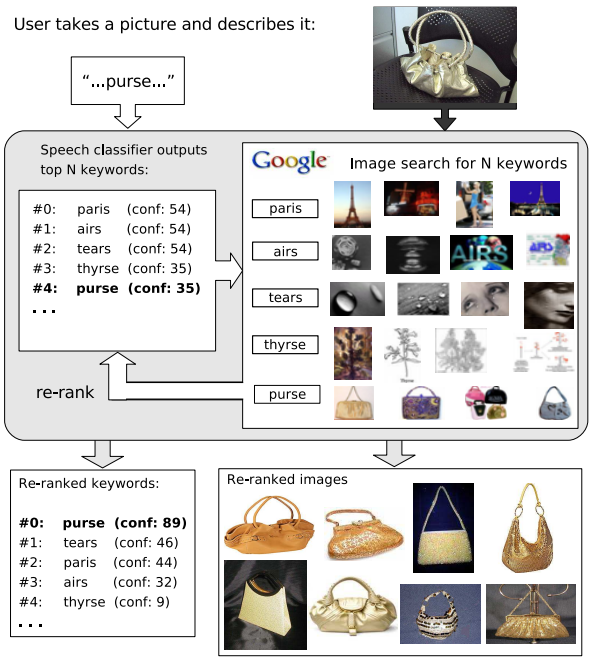

Figure 1: Concept figure illustrating adaptive multimodal object recognition. The system detects the user describing an object, e.g. "purse". The image is used to rank all images in an offline database (or online image search, as depicted above) that correspond to the top N keywords output by the speech classifier. The image match scores are then used to re-rank the kewords, and to output a filtered subset of images for the top keyword.

adaptation examples, the goal is to build the optimal visual category classifiers for that particular environment. Speech and language models can also be adapted to a particular speaker, although here we focus more on adaptation of image models.

Another form of adaptation is to learn out-of-vocabulary objects, i.e. objects whose images and/or referring words are not contained in the offline databases. In fact, learning out-of-vocabulary image models is necessary for a usable system, because current image databases cover very few object categories (on the order of a hundred), whereas there are far more objects that can be described in the English language (on the order of thousands). This can be acheived by exploiting weakly labeled images available on the web. Figure 1 illustrated this scenario in the context of the proposed system. Suppose the object "purse" is not in the visual vocabulary. Instead of using a labeled image database, we perform a keyword search in a large repository of weakly labeled (tagged) images, such as that provided by an online search engine. After disambiguating the correct keyword ("purse") based on both the user's speech and image data, we can build a visual model from the returned images. However, rather than keeping all the raw image results for "purse", which can be very noisy and contain images without the object, we obtain a subset by keeping those images that best match the input user image of the object, according to some distance metric (the "re-ranked images" in the figure). The visual model for the new object is then trained on these images.

Multimodal interaction using speech and gesture dates back to Bolt's Put-That-There system [1]. Since that pioneering work, there have been a number of projects on virtual and augmented-reality interaction combining multiple modalities for reference resolution. For example, Kaiser, et. al. [5] use mutual disambiguation of gesture and speech modalities to interpret which object the user is referring to in an immersive virtual environment. Our method is complimentary to these approaches, as it allows multimodal reference to objects in real environments, where, unlike in the virtual reality and game environments, the identity of surrounding objects is unknown and must be determined based on visual appearance.

Haasch, et. al. [4] describe a robotic home tour system called BIRON that can learn about simple objects by interacting with a human. The robot has many capabilities, including navigation, recognizing intent-to-speak, person tracking, automatic speech recognition, dialogue management, pointing gesture recognition, and simple object detection. Interactive object learning works as follows: the user points to an object and describes what it is (e.g., "this is my cup"). The system selects a region of the image based on the recognized pointing gesture and simple salient visual feature extraction, and binds that region to the object-referring word. Object detection is performed by matching previously learned object images to the new image using cross-correlation. The system does not use pre-existing visual models to determine the object category, but rather assumes that the dialogue component has provided it with the correct words.

The idea of disambiguating which object the user is referring to using speech and image recognition appeared in [7], where the authors describe a visually-grounded spoken language understanding system. An embodied robot is situated on top of a table with several solid-colored blocks placed in front of it on a green tablecloth. The robot learns by pointing to one of the blocks, prompting the user to provide a verbal description of the object, for example: "horizontal blue rectangle". The paired visual observations and transcribed words are used to learn concepts like the meaning of "blue", "above", "square". The key difference between this work and [7] again is that we focus on a realistic object categorization task, and on using prior visual models and weakly labeled databases. There has been some recent interest in using weakly supervised cross-modal learning for object recognition. For example, Fergus et al. [2] learn object categories from images obtained using a Google search for given text keywords.

2. DISAMBIGUATION

Our goal is to build a system that can learn to recognize objects in a situated environment with some help from the human user. We assume that there are no labeled examples of actual objects from the environment; the only labeled dataset available to the algorithm is a generic offline database, or images obtained by online search.

Two sub-problems arise in labeling new data: 1) how to locate the object the user is describing, and 2) how to obtain the correct object label from the speech and image data. The former can be difficult in general, as the object may not even be in the camera's view. To simplify the problem for now, we ask the user to take a picture of the object while describing it, which results in an image-waveform pair. The

second sub-problem is what we refer to as disambiguation. There are many cases in the literature where multimodal fusion helps recognition (e.g. [6], [5]). Although visual object *category* recognition is a well-studied problem, to the best of our knowledge, it has not been combined with speech-based category recognition. The success of such fusion will depend on whether there is enough complimentarity in the image and speech signals. Furthermore, it will depend on whether current image-based classifiers are good enough to improve on the speech results. Next, we present some experiments which show that it is, indeed, possible, using classic fusion algorithms, to benefit from fusion on this task.

Assume a fixed set of C categories, and a set W of nouns (or compound nouns), where W_k corresponds to the name of the kth object category, $k = 1, ..., C$. The inputs to the disambiguation algorithm consist of a visual observation x_1, derived from the image containing the object of category k, and the acoustic observation x_2, derived from the speech waveform corresponding to W_k. For the sake of this experiment, we assume that the user always says the same name for an object category (e.g., "car" and not "automobile".) A simple extension to multiple object names would involve mapping each category to a list of synonyms using a dictionary or an ontology such as WordNet.

We implement the disambiguation algorithm based on decision-level fusion of the outputs of the visual and speech category classifiers. Classifier decisions can be in the form of the class label k, posterior probabilities $p(c = k|x_i)$, or a ranked list of the top N hypotheses. There are several methods for fusing multiple classifiers at the decision level, such as letting the classifiers vote on the best class. We propose to use the probabilistic method of combining the posterior class probabilities output by each classifier. We investigate two combination rules. The first one, the weighted mean rule, is specified as:

$$p(c|x_1, ..., x_m) = \sum_{i=1}^{m} p(c|x_i)\lambda_i, \quad (1)$$

where m is the number of modalities, and the weights λ_i sum to 1 and indicate the "reliability" of each modality. This rule can be thought of as a mixture of experts. The second rule is the weighted version of the product rule,

$$p(c|x_1, ..., x_m) = \prod_{i=1}^{m} p(c|x_i)^{\lambda_i} \quad (2)$$

which assumes that the observations are independent given the class, which is a valid assumption in our case. The weights are estimated experimentally by enumerating a range of values and choosing the one that gives the best performance on a held-out dataset. Using one of the above combination rules, we compute new probabilities for all categories, and pick the one with the maximum score as the final category output by the classifier.

Figure 2 shows the results of fusion on a 101-category dataset consisting of image-waveform pairs. The waveforms consisted of recordings of six speakers saying the names of objects depicted in images from the *Caltech*101 database. The image-based object classifier was implemented using a support vector machine operating on a histogram over visual words [3]. The speech recognizer's vocabulary was limited to the set of phrases defined by W. To simulate a more realistic speech task, we added "cocktail party" noise to the original waveforms, using increasingly lower signal-to-noise ratios (SNRs): 10db, 4db, 0db, and -4db. The plot shows the gains that each combination rule achieved over the single modality classifiers. The mean rule (red line) does slightly better than the product rule (green line) on a number of noise conditions, and significantly better than the either speech or vision alone on all conditions. Further details of the experiment may be found in [8].

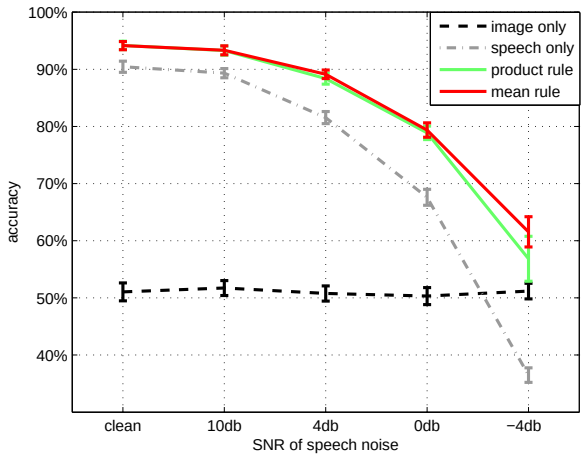

Figure 2: Test results of disambiguation across speech noise conditions. The Y-axis shows the percent of the test samples classified correctly; the X-axis shows the SNR of the noise condition; chance performance is around 1%.

Although these results show promise for multimodal object recognition, many issues remain to be solved. One such issue is the mapping the object-referring words to the object category label. Is there a difference between "cup" and "paper cup"? Should we treat both instances as the same *basic category* object, or as two distinct categories? Also, words can have more than one meaning, for example, "cup" can mean a container for drinking or a sports trophy, as in "the Stanley Cup". Therefore, word sense disambiguation may be a crucial step in determining the correct label, and, in fact, multiple word senses can potentially be disambiguated based on the image.

3. ADAPTATION

Two potential problems arise when using visual object classifiers trained on offline data in a new environment: 1) training and test image mismatch, and 2) insufficient object inventory. The first is the problem of generalizing from the kinds of objects contained in the database to the ones in the particular home or office where the system is deployed. Most existing databases, for example, *Caltech 101*, are made up of images mined from internet search engines (see top two rows of Figure 3). Unfortunately, web images are not representative of the types of pictures we might take in our physical surroundings. They are often shot by professional photographers, or at least aim to be aesthetically pleasing. They rarely have blurring or occlusion, and the objects are typically centered in the image and have canonical pose. Other features of web images include studio lighting, blank back-

Figure 3: Top row: first three images returned from Google image search for keyword "laptop"; middle row: sample laptop category images in Caltech 101 database; bottom row: images of laptops taken in a typical home.

grounds, and bright colors. On the other hand, a robot in the real world would probably encounter images with poor lighting, blurring, and random pose (see bottom row of Figure 3). The location, background, and appearance of objects are more likely to be an important cue for recognition, because the robot will encounter the same objects repeatedly.

The problem of insufficient object inventory is that the object database will most likely not include all of the types of objects present in the environment. In fact, even if the offline database covers tens of thousands of categories, there will always be category *instances* that are not in the database. People's faces are a good example of this. The offline classifier may be able to distinguish a male face from a female one, but will not be able to tell "Bob" from "Mike" unless those individuals are added to the database.

Adaptation provides a solution to both problems. Two adaptation methods can be applied: unsupervised and supervised. In unsupervised adaptation, existing models are tuned to the new environment using only unlabeled images. One form of unsupervised adaptation is image normalization, such as color and brightness histogram equalization, blurring, and noise correction. However, in addition to differences in imaging conditions, there are differences in the nature of the tasks. The offline databases are typically designed to benchmark object recognition algorithms, and therefore have a lot of variation within each category. In a real world environment, the variation within a category is not going to be as large. For example, there are unlikely to be more than a few different types of phones in any given office. Therefore, the intra-class variation for a category such as "phone" is going to come not from many different instances, but rather from different views of the same few instances.

In supervised adaptation, we augment the offline training data with unlabeled data of the user describing objects in her immediate environment. The pairs of images and spoken descriptions provide additional labeled object data that can be used to adapt the offline object classifier, and to learn out-of-vocabulary objects. The adaptation technique will depend on the variation between the original and target environments, and on the classifier used by the system. This work is ongoing, and we plan to explore several approaches, including feature space adaptation and re-weighting of training examples.

4. CONCLUSION

We have described an approach to situated object categorization that makes use of prior models of both speech and images. The system will learn gradually, using input from the user to adapt its existing models, and to learn new ones. We showed experimentally that it may be advantageous for HCI systems to use both channels to recognize object references, as opposed to the conventional approach of relying only on speech or only on image recognition, when both are available. We regard this work as the first step towards multimodal object category recognition. We plan to continue this line of research, extending the model to handle multiple words per category, and to extract possible object-referring words from natural dialogue. Although we have focused on an assistant robot application, the methods described here could be applied to automatic speech-based photo tagging, personal video tagging on a mobile phone, and other situations where the user provides spoken descriptions of objects in his or her environment.

5. REFERENCES

[1] R. Bolt: "Put-that-there": Voice and gesture at the graphics interface. In Proceedings of the 7th Annual Conference on Computer Graphics and interactive Techniques. SIGGRAPH '80. ACM Press, New York, NY, 262-270, 1980.

[2] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman: Learning Object Categories from Google's Image Search. Proc. of the 10th Inter. Conf. on Computer Vision, ICCV 2005.

[3] K. Grauman and T. Darrell: The Pyramid Match Kernel: Discriminative Classification with Sets of Image Features. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Beijing, China, October 2005. Software available at: http://people.csail.mit.edu/jjl/libpmk/

[4] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer: A multi-modal object attention system for a mobile robot, Intelligent Robots and Systems. 2005.

[5] E. Kaiser, Olwal, A., McGee, D., Benko, H., Corradini, A., Li, X., Cohen, P., and Feiner, S.: Mutual disambiguation of 3D multimodal interaction in augmented and virtual reality. In Proceedings of the 5th International Conference on Multimodal Interfaces (ICMI). 2003.

[6] G. Potamianos, C. Neti, G. Gravier, A. Garg, and A. Senior: Recent Advances in the Automatic Recognition of Audio-Visual Speech, in Proc. IEEE. 2003.

[7] D. Roy, P. Gorniak, N. Mukherjee, and J. Juster: A Trainable Spoken Language Understanding System for Visual Object Selection. In Proceedings of the International Conference of Spoken Language Processing. 2002.
[8] Reference omitted for blind review.