Semi-supervised Object Recognition Using Images and Speech

Ph.D. Thesis Proposal by Kate Saenko

Computer Science and Artificial Intelligence Laboratory
MIT, Cambridge, MA
saenko@mit.edu

Thesis Supervisor: Trevor Darrell

April 13, 2007

Abstract

Object recognition is an important part of human-robot interaction (HRI) 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 propose to develop a semi-supervised algorithm for learning about physical objects found in a situated environment based on visual and spoken input provided by the user. The algorithm would operate on generic databases of labeled object images and transcribed speech data, plus unlabeled video 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. Humans use multiple modalities to understand which object category is being referred to, simultaneously interpreting gesture, speech and visual appearance, and HRI systems should be able to do the same. 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. An additional interesting extension of the proposed approach is to learn out-of-vocabulary objects, i.e. objects whose images and referring words are not in the generic labeled databases. This would be acheived by exploiting unlabeled images available on the web, matching the reference image of the unknown object to images returned by the image search for the candidate category words.

Contents

1 Introduction

Automatic scene understanding, or the ability to categorize places and objects in the immediate environment, is important for many HRI applications, including mobile robotic assistants for the elderly and the disabled. *Category*-level recognition allows the system to recognize a class of objects, as opposed to just single instances, and is particularly useful. One approach to automatic scene understanding is through *image*-based recognition, which involves training a classifier for each scene or object category offline, using manually labeled images. However, to date, image-based category recognition has only reached a fraction of human performance, especially in terms of the variety of recognized categories, partly due to the lack of labeled data. Accurate and efficient off-the-shelf recognizers are only available for a handful of objects, such as faces and cars. Thus, to enable an assistant robot, or a similar system, to accurately recognize objects in the environment, the user currently would have to collect and manually annotate sample images of those objects.

Alternatively, a robot can learn about its surroundings from interactions with the user. One example is the "home tour" scenario [9], where the user points to objects around the room and describes them verbally, e.g., "this is my pen". Another example is the visually-grounded language acquisition system of [14], where the robot prompts the user to describe the objects located on a table by pointing to them with a laser pointer. However, neither of these "show-and-tell" systems use any prior knowledge of object category appearance. Rather, they perform speech-only object labeling, using the output of a speech recognizer to determine object-referring words, and then using those words directly as object labels. Thus, if the spoken description is misrecognized, an incorrect object label may be assigned to the input image (e.g., "pan", instead of "pen".) The other disadvantage of this approach is that the system is limited to visually simple objects and backgrounds, since it relies on regions of solid color to segment a candidate object from its background. Also, it learns only from the examples provided by the user. It therefore only knows about object *instances* that the user has pointed out. This places a burden on the user to show the robot many different objects belonging to the same category, in order for it to generalize to unseen objects.

In this work, our goal is to enable human-computer interaction systems to recognize a variety of object categories in realistic environments without requiring manual annotation of each category by the user. We propose a new approach, combining speech and visual object category recognition. The approach consists of two parts: *disambiguation* and *adaptation.* Disambiguation means that, instead of relying completely on one modality, we will use generic visual object classifiers to help the speech recognizer obtain the correct object label. The generic visual classifiers are supervised classifiers trained on an offline image database that is not tailored

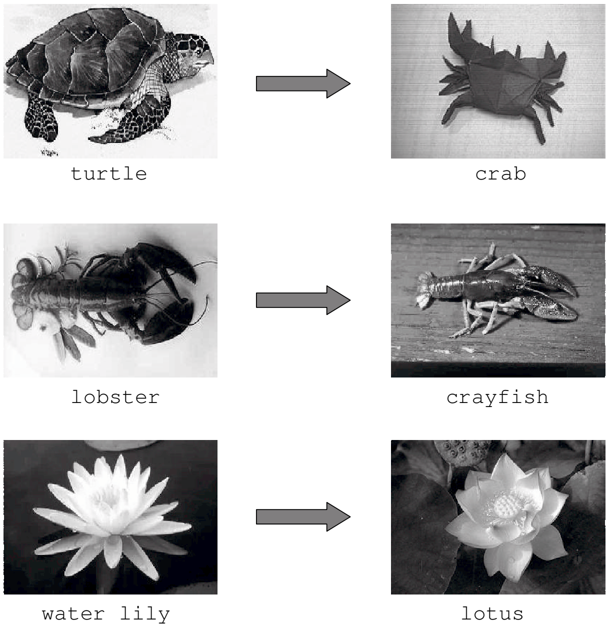

Figure 1: Examples of the most visually confusable categories in our dataset (see Section 5 for a description of the experiments). The image-based classifier most often misclassified the category on the left as the category on the right.

to the specific environment the system may be deployed in. Although we do not expect these classifiers to be highly accurate, our hope is that they will be useful in constraining the recognition of *acoustically* similar object-referring words. For example, if the ASR component returns “my nose” as the most likely word sequence, and “my notes” as the second most likely one, the generic visual classifier might be able to rule out the first choice based on the corresponding image. The same intuition applies in the other direction, with speech disambiguating confusable visual categories. For example, Figure 1 shows the categories that the visual classifier confused the most in our experiments in Section 5. Since the word pairs are not very acoustically similar, the speech-based classifier is able to correct the visual errors.

In the adaptation step, we adapt object classifiers trained on generic databases to specific objects found in the environment, using new labels obtained in the disambiguation step as adaptation data. This approach can be thought of as a form of knowledge transfer from one scene (the database) to another scene (the environment). The goal of adaptation is, given a labeled generic database and a small number of labeled adaptation examples, to build the optimal visual category classifiers for that particular environment. Speech and language models can also be adapted to the particular speaker. Since image databases are limited in the number of categories, another goal of adaptation can be to learn out-of-vocabulary objects, i.e. objects whose images and referring words are not in the generic labeled databases. This can be acheived by exploiting unlabeled images available on the web. For example, we can match the reference image of the unknown object to a subset of images returned by an image search for the top most likely words for that object.

In the next section, we will describe a few potential applications which motivate this research. Then, in Section 3, we describe the proposed approach. Section 4 talks about databases that may be useful to us, and Section 5 describes preliminary and future planned experiments. Section 6 reviews related work, and Section 7 gives a proposed timeline for completing the proposed work.

2 Potential Applications

Before discussing the proposed approach in more detail, we describe a few specific applications. The first application is a robotic "companion" which assists people in their home or office (see Figure 2). Mobile robotic assistants have the potential to provide valuable services, especially for the elderly and the disabled [12]. The vision module of such a robot must identify objects in its environment, and respond to user requests, such as "pick up the newspaper off the floor". Ideally, this should happen automatically, without any additional training required by the user. However, that is not feasible in practice, as even a human companion would need to be taught what certain objects are and where they are located in the home. Presumably, the most natural way for a human to "teach" a robot is by interacting with it as if it were another human. For example, the user should be able to give the robot companion a tour of the home, pointing out things and rooms, while the robot asks questions and confirms its recognition hypotheses.

Other applications do not involve an embodied robot, nor conversational interaction. The user can simply describe things while recording them with a camera, and the algorithm can use the available narration to learn about the objects in the video. One advantage of this approach is not having to wait for the robot to navigate, recognize gestures and objects, and generate responses, which may all be time-consuming, at least for the current state-of-the-art systems. Another advantage is that the user can point the camera at the object of interest, which may be easier than having the robot recognize pointing gestures. Such a system can be used for surveillance, as an automatic guide, a language tutor, or as assistive technology for the visually impared. One example of a non-embodied application is a passive capture system for documenting a user's experiense, which has a wearable camera

Figure 2: Assistant robots: Asimo humanoid robot developed by Honda (left), Care-o-bot assistant robot developed by the Fraunhofer Institute for Manufacturing Engineering and Automation (right).

that selectively records images and can serve as a memory aid (e.g. “where did I park my car?”)

In this research, we will focus on the non-embodied conversational system scenario, however, in principle, the algorithms should be transferrable to the embodied case. We will assume that the user will either wear a head-mounted camera, or point a hand-held camera at the entity of interest. A (very) preliminary experiment involving one user on how one might give a home tour in this setting showed that it was most natural for the user to point a hand-held camera at objects, one by one, and speak the name of the object while keeping it in view for a few seconds. For example: ”this is my office... this is the couch... chair... cup... laptop...,” etc. Typically, the user took a close-up of the object while talking about it. The user found it easier to take a close-up shot of the object, rather than take a far-away shot that may include other objects, and then use the free hand to point to its exact location, or verbally describe the exact object in terms of color and other properties. However, the resulting images were not just of the object, due to background clutter and to some objects being too large or too small (see Figure 5). We plan on collecting more data in this way, with multiple users describing objects in a specific environment (see Section 5).

3 Learning from User-Narrated Video

In this section, we discuss the two problems that we propose to address, those of disambiguation and adaptation, in more detail. Again, 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 the specific objects in the environment; the only labeled dataset available to the algorithm is a generic offline database.

3.1 Disambiguation

Adapting existing object classifiers to new environments seems like a promising approach. However, in order to obtain labeled data for the adaptation step, we must first solve the following problems:

- how to locate the object the user is describing, and
- how to obtain the correct object label from the speech and image data.

To address the first problem, in the initial stages of this research, we will ask the user to take a picture of the object during training, and will assume that the object will be more or less centered and occupy most of the image. In testing, we will only evaluate the classifier, which, given a location and size estimate, will output the probability that the object belongs to a particular class. The size and location can either be exhaustively searched in the image, or can come from a global scene context model. For now, we will use manually specified location and size parameters for testing our classifiers.

The second problem is what we refer to as disambiguation. In disambiguation, the task is to use the spoken description of an object and the image of the object to produce the object category label. At first glance, this task may seem straightforward. We can use the speech recognizer to generate a transcript of what the user said, locate the word referring to the object (eg. "cup" in "this is my cup"), and use that word to label the image. However, there are several reasons why this may not work. The first is that the speech recognizer might make errors. State-of-the-art speech recognition systems can achieve high word recognition rates, depending on the vocabulary and the constraints imposed by the language model. However, recognition of isolated object descriptions with a large vocabulary (in the tens of thousands of words) has not been explored in the speech recognition literature. One solution might be to lower the complexity of the recognition task by restricting the user to a fixed-size vocabulary. For example, to describe a dog, the user would only be able to say the word “dog”, and not “doggie”, “puppy”, or “terrier”. However, that would make the system very non-transparent and difficult to use.

There are many cases in the human-computer interaction literature where multimodal fusion helps recognition (e.g. [13], [8]). 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. In the experimental section, we will use real images, as well as speech waveforms from users describing objects depicted in those images, to see whether there is complementary information in the two channels. We propose a fusion algorithm based on probabilistic combination of the speech and image classifier outputs. We show that it is feasible, using state-of-the-art recognition methods, to benefit from fusion on this task. The current implementation is limited to recognizing about one hundred objects, because that is the number of categories in the labeled image database. In the future, we will explore extensions to allow arbitrary vocabularies and categories.

For now, we 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 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. In the initial experiments, we assume that the user always uses the same name for an object category (e.g., "car" and not "automobile".) In the future, we will address an extension to multiple object names. A simple extension would involve mapping each category to a list of synonyms using a dictionary or an ontology such as WordNet.

The disambiguation algorithm consists of decision-level fusion of the outputs of the visual and speech category classifiers. Currently, the speech classifier is a general-purpose recognizer, but its vocabulary is limited to the set of phrases defined by W. Decision-level fusion means that, rather than fusing information at the observation level and training a new classifier on the fused features $x = \{x_1, x_2\}$, the observations are kept separate and the decision of the visual-only classifier, $f_1(x_1)$, is fused with the decision of the speech-only classifier, $f_2(x_2)$. 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:

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

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,

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

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. 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.

Our visual classifier is a multi-class SVM, which returns margin scores rather than probabilities. To obtain posterior probabilities $p(c = k|x_2)$ from decision values, a logistic function is trained using cross-validation on the training set. Further details can be found in [1].

One issue in disambiguation that we need to solve is mapping the object-referring words to the object category label. For example, what is the difference between "cup" and "paper cup"? Should we treat both instances as the same *basic category* object, or as two distinct categories? If the user uses both phrases to refer to the same object, should the system return the same object category label? Also, words can have more than one meaning, for example, "cup" can mean a liquid 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.

Our intuition is that using generic classifiers trained on existing labeled image data can potentially help solve some of these issues. For example, disambiguation could help reduce ASR errors, if acoustically ambiguous words such as "cat" and "hat" correspond to non-ambiguous visual categories. Also, multiple word senses can potentially be disambiguated based on the image. However, there are still many open questions surrounding this line of research. One such question is whether current object recognition technology is up the task. In Section 5, we describe the set of experiments we hope will help answer these questions.

3.2 Adaptation

Two potential problems can arise when using visual object classifiers trained on offline data in a new environment:

- training and test image mismatch
- insufficient object inventory

The first is the problem of generalizing from the kinds of objects 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 Figure 3). Unfortunately, web images are not representative of the types of images we see in our physical surroundings everyday. They are typically shot by professional photographers, or at least amatures trying to create aesthetically pleasing results. Thus, these images usually have no 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 backgrounds, and bright colors. On the other hand, a robot in the real world would probably encounter images with poor lighting, blurring, and random pose. Also, the location, background, and color of objects are more likely to be an important cue for recognition, because the robot will encounter the same objects repeatedly. The *Labelme* database is different from *Caltech 101* in that it contains some real-world images, not mined from the web but shot by people in specific environments. For example, Figure 4 shows sample objects belonging to the "lamp" and "laptop" categories in the Labelme database, including the manual outlines of the objects. Although these show objects in natural environments, they are still quite different from images in Figure 5, which shows objects of the same categories found in a specific home.

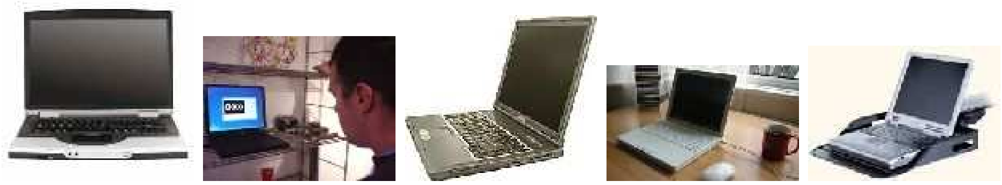

Figure 3: Sample laptop category images in Caltech 101 database.

The second problem is that the object database will most likely not include all of the types of objects present in the environment. In fact, even if our 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 "Adam" unless those individuals are added to the database.

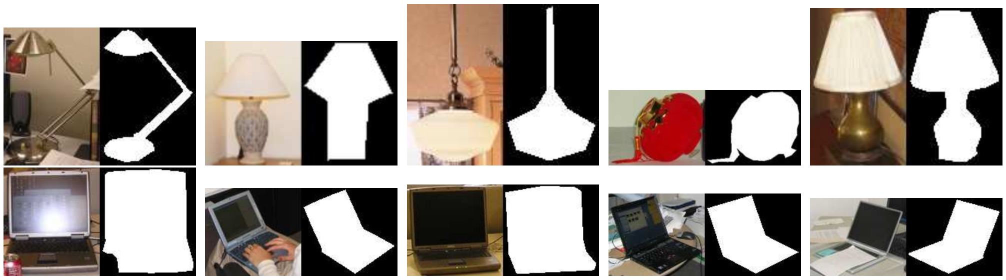

Figure 4: Several of the results of queries for "lamp" and "laptop" in the *Labelme* database.

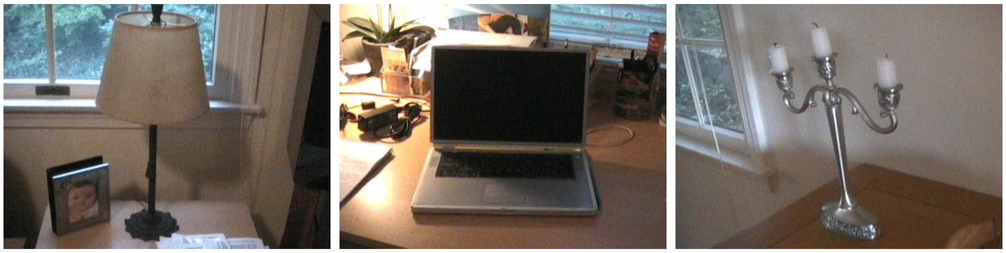

Figure 5: Frames corresponding to words "lamp", "laptop" and "candlestick" in a sample user-narrated video of a home.

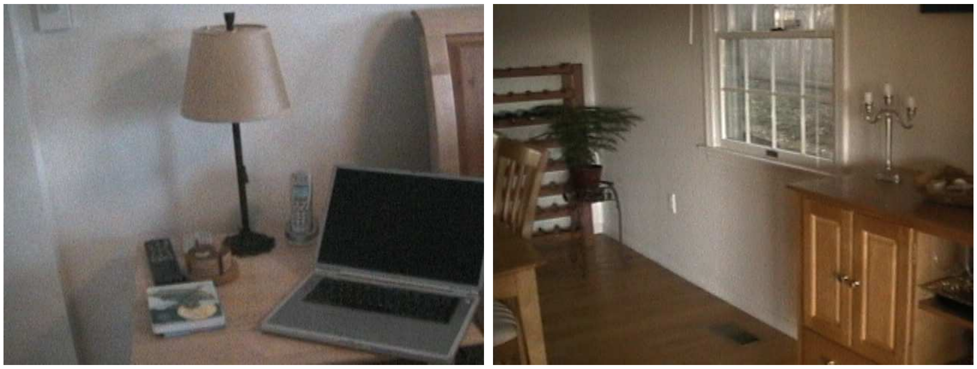

Figure 6: Novel test images of lamp and laptop (in a new scene) and candle (in the same scene).

The solution to the problem of data mismatch is to adapt the classifiers. One way to do this is by unsupervised adaptation, tuning existing classifiers 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. But, in addition to differences in imaging conditions, the tasks are also of different nature. 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 variety within a category is not going to be as large for each object category. For example, there are unlikely to be more than a few different types of phones in any given office. That means that the intra-class variation for the category "phone" is going to come not from many different instances, but rather from different views of the same few instances.

We propose to adapt classifiers to the new task in a semi-supervised manner, by augmenting 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 exact adaptation technique we will use will depend on the variation between the original and target environments, and on the classifier used by the system. We plan to explore several approaches, including feature space adaptation and re-weighting of training examples. Note that the adaptation step can also apply to the speech modality. While speaker adaptation is a well-explored problem in ASR, it has not been studied in this particular context.

4 Existing Speech and Image Datasets

This section briefly describes several existing datasets that may be of use to us for training generic object models and speech recognizers. Most publicly available image databases suitable for category-level recognition contain either cars or faces, and very few other object categories. The exceptions include the *PASCAL*, *LabelMe*, *Caltech101*, *ESP* and *Peekaboom* databases, described in [16].

4.1 Caltech101

For the disambiguation experiments described in Section 5, we chose to use the *Caltech101* database, because it contains a large variety of categories, and because it is a standard benchmark in the object recognition field. The database has a total of 101 categories, with about about 50 images per category. Although the categories are challenging, the task is made somewhat easier by the fact that most images have little or no clutter, and the objects tend to be centered in each image, presented in a stereotypical pose. Sample images from each of the 101 categories are shown in Figure 7.

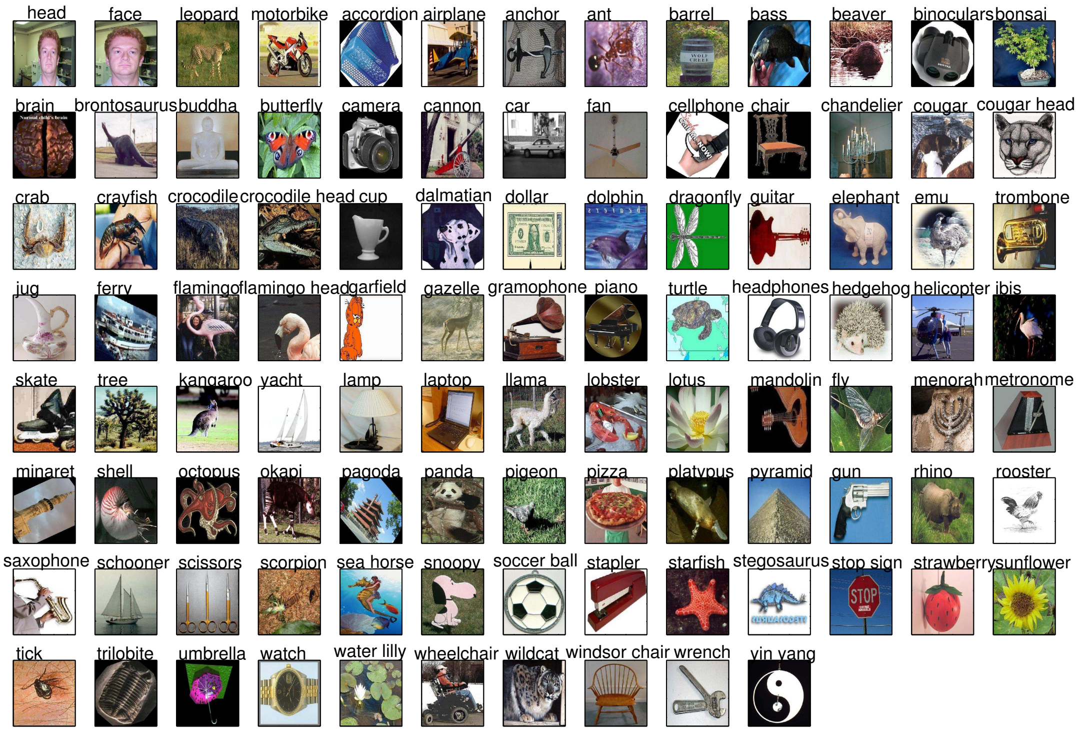

Figure 7: Sample images from the *Caltech101* database. The category name used in our experiments is shown at the top of each image.

4.2 LabelMe

LabelMe is a freely available richly annotated image database. It consists mainly of office and street scenes. There are currently a total 114,673 objects annotated in the database. Labels include a text string and the outline of the object. Samples from the database are shown in Figure 4.

4.3 Phonebook

PhoneBook is a phonetically-rich, isolated-word, telephone-speech database. The goal of PhoneBook is to serve as a large database of American English word utterances incorporating all phonemes in as many segmental/stress contexts as are likely to produce coarticulatory variations, while also spanning a variety of talkers and telephone transmission characteristics. The core section of PhoneBook consists of a total of 93,667 isolated-word utterances, totalling 23 hours of speech. This breaks down to 7,979 distinct words, each said by an average of 11.7 talkers, with 1,358 talkers each saying up to 75 words. Talkers were adult native speakers of American English chosen to be demographically representative of the U.S.

4.4 The Human Speechome Project

The Human Speechome project [15] is a data collection effort on an unprecedented scale designed to record every waking moment of a child's first few years of life, using microphones and cameras mounted in the ceiling of each room in the child's home. The goal of the project is to train and evaluate compuational models of visually grounded language learning, based on everything the child sees and hears. The dataset is projected to grow to 142,000 hours of video and 196,000 hours of audio by the end of the three year effort. Although this might be a useful audio-visual dataset for our purposes, it is not clear if and when it will be made publically available.

5 Planned Initial Experiments and Data Collection

Before we embark upon building an integrated system, we plan to carry out some preliminary experiments. Their purpose is to serve as a proof-of-concept, to explore the various issues that might arise in a controlled setting, and to guide the development of the final system. In this section, we will describe the first set of planned experiments, which will explore visual object recognition accuracy, speech recognition accuracy, and speech and image data collection. We will also present the results of our preliminary experiment on disambiguation of object categories using the *Caltech101* database.

5.1 Planned Object Recognition Experiments

The object recognition experiments will proceed as follows. An object classifier will be trained on an offline database. We plan to use a state-of-the-art algorithm such as the the SVM-based method of Grauman and Darrell [3]. To our knowledge, there are no published classification results on the *LabelMe* database, so part of the experiment is to get a sense of what accuracy we can expect on that database. The other goal is to estimate the performance on the target environment images.

For this and the following experiments, we need to choose a training database from the several image datasets freely available to the research community. We also need to collect a database of test images from a target environment. For now, we will assume that each image in both the train and test datasets contains only one object, and that the object occupies the majority of the image. That means that we are not dealing with the problem of finding the object in the test image, just assigning the right label given the object's size and location. The following describes the general setup of the experiment:

Choose Labeled Database: Unfortunately, most public databases have mainly images of cars and faces, and very few other object categories. Exceptions are *LabelMe*, *Caltech 101 and 256*, *Google Image* [6] and *Peekaboom* [11]. The latter two are not technically databases, but rather search engines, but one could collect a database by searching them for object categories.

Select Sample Objects: For this experiment, we will assume that a trained classifier exists for the exact test object category, i.e. if a test image is of a "laptop", there should be a "laptop" category in the training database. This means that the test categories should be a subset of the training categories. Examples of possible objects are: keyboard, mouse, chair, mug. Altogether we aim to have about 20-30 categories with about 100 images per category.

Collect Test Images: We will collect test images in an office environment, specifically, in a graduate student office and/or lab space. Objects will be placed in typical places around the room, i.e. a cup will be placed on a table rather than the floor. No effort will be made to match the images in the training database. Volunteer subjects will be asked to take several pictures of each object. The objects will be placed in different location from one session to the next to create different backgrounds, viewpoints and illumination.

5.2 Planned Speech Recognition Experiments

Before collecting new speech data from real users, we plan to carry out several simulated experiments. The main goal of the simulated speech recognition experiment is to decide on a reasonable vocabulary of object-referring words, and to establish the approximate accuracy of the speech recognizer using the chosen vocabulary. For this experiment, we will first choose an existing speech corpus containing isolated words that describe physical objects. We will then use the corpus to train and test an isolated word recognizer. The following is a list of questions the experiment will address:

- Which (multi-speaker) speech corpus should we use to train and test the recognizer? Phonebook is one option.
- What should the vocabulary be? There are several options: all English words, all nouns, all nouns that correspond to a physical entity (based on the WordNet definition), or a more restricted set of words, such as the list of all products sold on an online office supply website.
- How accurate is the resulting speech recognizer?
- What is the length of the N-best list guaranteed to contain the correct hypothesis?

The above should give us a rough idea of what to expect from the speech recognition, and provide a baseline for the following experiments.

5.3 Collecting Spoken Descriptions from Real Users

The goal of the next experiment is to collect speech data of real users describing familiar objects in a specific environment. We want to see how users natually describe typical objects in familiar surroundings. This data will be used in the audio-visual disambiguation experiment.

The data collection will proceed as follows. Each subject will be presented with images of objects from the test data collected in the first experiment. For each image, they will be asked to say what the object is. To achieve a more natual interaction style, the subjects will be told to imagine describing the objects for the sake of a friend, who will later listen to the descriptions. To make sure that subjects are describing objects familiar to them, they will be asked to briefly familiarize themselves with the images prior to the experiment. The advantage of using existing images rather than asking the user to take pictures is that we can more easily control which environment and which objects are used. We aim to have around ten users, each describing 5 images per category, 20 categories. This would amount to 50 utterances total per category.

We hope for this experiment to address the following questions:

- What is the WER using the speech recognizer on this data?
- What words did the subjects use? Did they form complete sentences?
- Can the object-referring label be predicted using the list of synonyms?
- What is the percentage of out-of-vocabulary words?

5.4 Results of a Simulated Disambiguation Experiment

We have carried out the following experiments, with real images and speech from users describing objects depicted in those images, to see whether there is complementary information in the two channels. If so, then fusing the classifiers should give us better recognition performance than using either one in isolation. We are not aware of any publicly available databases that contain paired images and spoken descriptions. For these experiments, we used a subset of a standard image-only object category database, and augmented the images with speech by asking subjects to view each image and speak the name of the object category. We evaluate the proposed fusion algorithm and compare the mean and the product rules.

To train the image-based classifier, we use a standard training set of the *Caltech101* database, consisting of the first 15 images from each category. We also select a test set consisting of the next 12 images from each category, for a total of 1212 test samples. Each image in the test set was paired with a waveform of a subject speaking the name W_k of the corresponding category. All experiments were done averaging the performance over 20 trials of randomly selecting subsets of 50% of the test data.

We chose the set W based on the words provided with the image database, changing a few of the names to more familiar words. For example, instead of "gerenuk" we used the word "gazelle". The exact W_k is shown in Figure 7. A total of 6 subjects participated in the data collection, 4 male and 2 female, all native speakers of American English. Each subject was presented in turn with 2 images from each category in the image test set, and asked to say the object name, resulting in 12 utterances for each category. The reason that the images were shown, as opposed to just prompting the subject with the text string, is that some names have ambiguous spellings (e.g., the spelling "bass" refers to the fish, not the musial instrument), and also to make the experience more natural. Our goal is to simulate the scenario where the user speaks the name of the object and points to it, with the robot interpreting the gesture and thus obtaining an image of the object.

The speech data collection took place in a quiet office, on a laptop computer, using its built-in microphone. The nature of the category names in the *Caltech101* database, the controlled environment, and the small vocabulary makes this an easy speech recognition task. In realistic human-robot interaction scenarios, the environment can be noisy, interfering with speech recognition. Also, the category names for everyday objects are more common words (e.g. "pen" or "pan" instead of "trilobyte" or "mandolin") and the their vocabulary is much larger, resulting in more acoustic confusion. To simulate a more realistic speech task, we added "cocktail party" noise to the original waveforms, using increasingly lower signal-to-noise ratios: 10db, 4db, 0db, and -4db.

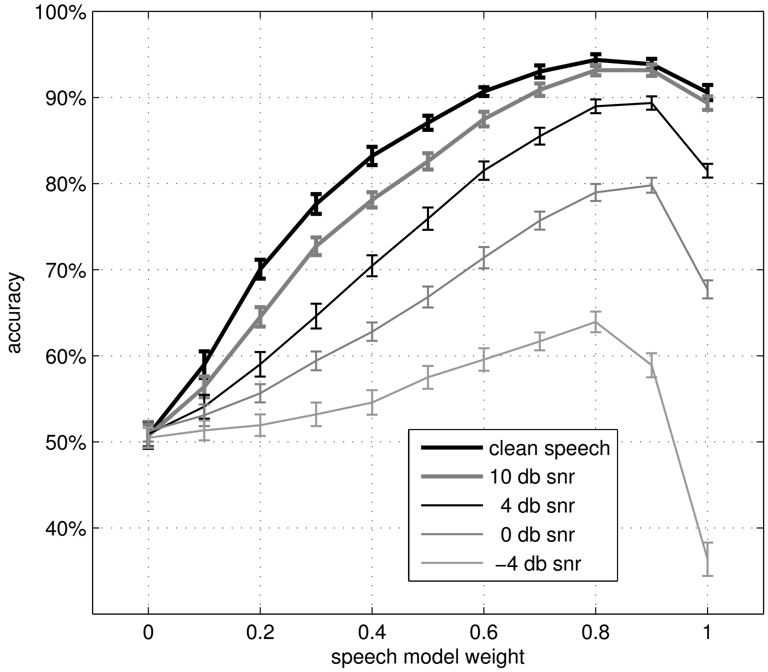

Figure 8: Multimodal object recognition using the mean rule. Each line represents the performance on a different level of acoustic noise. The y-axis shows the percent of the test samples classified correctly, the x-axis plots the speech weight used for the combined classifier.

For these expriments, we used the Nuance speech recognizer, a commercial, state-of-the-art, large-vocabulary speech recognizer. The recognizer returns an N-best list, i.e. a list of N most likely hypotheses $k = k_1, ..., k_N$, sorted by their confidence score. We use normalized confidence scores as an estimate of the posterior probability $p(c = k|x_1)$ in the combination rule. For values of k not in the N-best list, we set the probability to 0. The size of the N-best was set to 101, however, due to pruning, most lists were much shorter. The total 1-best accuracy on the entire test set obtained by the recognizer in the clean audio condition was 91.5%. The accuracy is measured as the percentage of waveforms assigned the correct category label. The N-best accuracy, i.e. the accuracy that would be obtained if we could choose the best hypothesis by hand from the N-best list, was 99.2%.

We use the method of [3] for image-based category classification. The algorithm first extracts a set of interest points from the image, and then performs vector quan-

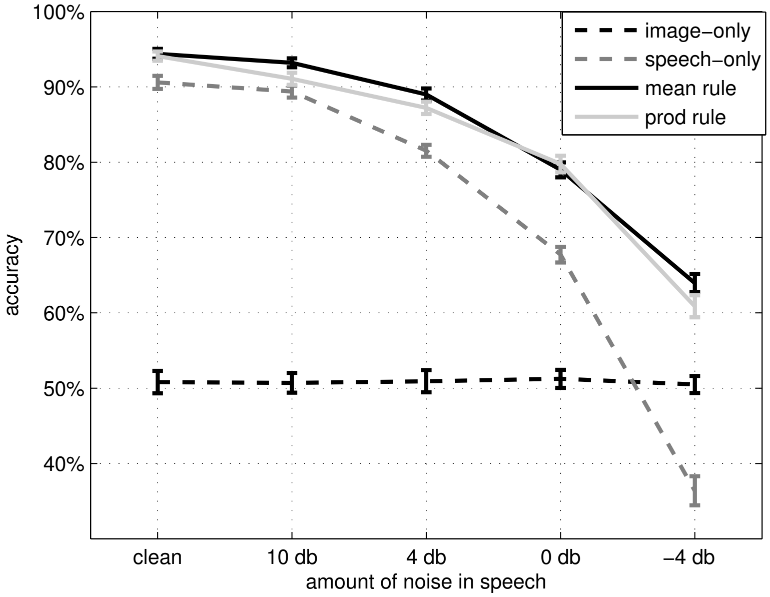

Figure 9: Absolute improvement over noise conditions, for speech weight = 0.8. The Y-axis shows the percent of the test data classified correctly, the X-axis the noise condition.

tization on the feature space [4]. Classification is done with a multi-class support vector machine (SVM) using the pyramid match kernel [3]. The implementation uses an all-vs-one SVM, with a total of C classifiers, each of which outputs posterior probabilities of each class given the test image. The classification accuracy obtained on the entire test set by the image-based classifier, measured as the percentage of correctly labeled images, was 50.7%. Note that it is possible to achieve better performance (58%) by using 30 training images [5].

Figure 8 shows the results of the fusion algorithm using the mean combination rule. Each line represents the performance on a different level of acoustic noise, with the top line being clean speech, and the bottom line the noisiest condition, -4db SNR. The x-axis plots the speech model weight λ_2, where $\lambda_1 + \lambda_2 = 1$. Thus, the leftmost point of each line is the average image-only accuracy, and the rightmost point is the speech-only accuracy. As expected, speech accuracy degrades with increasing noise. We can see that the fusion algorithm is able to do better than either single-modality classifier for some setting of the weights. The product combination rule gives similar performance to the sum rule, therefore we do not show the detailed results here.

We see from the results that the weighted combination rule is better than not having weights (i.e. setting each weight to 0.5). The optimal weight can be estimated automatically based on a held-out dataset not used in testing. Our results show that, for any noise condition, setting the speech model weight to 0.8 seems like the best choice, even when the speech modality performance is much worse than visual performance. The absolute gains in classification accuracy at that weight for each noise condition are plotted again for clarity in Figure 9. This figure shows that the mean rule does slightly better than the product rule on a number of noise conditions.

6 Related Work

In this section we review existing work in several related areas.

6.1 HCI-based Vision Systems

In the case of an embodied robotic companion, Haasch, et al. describe a robotic home tour system called BIRON that can learn about simple objects by interacting with a human. The system has many other 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. Object detection works by matching previously learned object images to the new image using cross-correlation. When it fails to identify a region, the robot can ask the user what color the object is to refine the search.

While the human-robot interaction is an impressive contribution of the above system, the object detection part could probably be improved. For example, in a realistic environment with clutter and non-uniform backgrounds, simply pointing at the object can be inaccurate. Also, the simple region extraction and correlation-based recognition methods will only work for simple objects, such as a uniformly-colored cup on a green table cloth. The challenge is to build a system capable of dealing with a realistic scenario in terms of clutter and variety of objects. Note that this work does not use pre-existing models of object, but rather learns them just from the examples provided by the user.

6.2 Object Recognition

There is a large body of work on object recognition in the computer vision literature, a comprehensive review is beyond the scope of this document. Here we only review several of the most recent publications, concentrating on object presence detection, where, given an image, the task is to determine if a particular object category is present, and object classification, where the task is to determine which one out of N categories is present in the image.

Murphy et. al. use a context-sensitive object presence detection method [10]. The overall image context gives the probability of the object being present in the image, which is used to correct the probability of detection based on the local image features. The authors show that the combination of experts based on local and global image features performs better than either expert alone. Our proposed method is somewhat similar to this, except that it is a combination of experts based on speech and image features.

The current two best-performing object classification methods on *Caltech 101* are the methods of Frome et. al. [2] and Zhang et. al. [17]. In [2], a nearest-neighbor classifier is used in combination with a perceptual distance function. This distance function is learned for each individual training image as a combination of distances between various visual features. The authors of [17] use a multi-class support vector machine (SVM) classifier with local interest point descriptors used as visual features. Both methods achieve around 66% average classification accuracy with 30 training images.

6.3 Semantic Concept Retrieval from Video

Hoogs, et al. [7] developed an automatic video annotation system that uses visual category recognition in combination with a WordNet-based ontology. Given a video of a news story segmented into a series of clips, the system outputs a list of recognized concepts for each clip, for example, “helicopter”, “lifeboat”, etc. To determine the correct concepts, the images are segmented into regions and each region is classified as one of several visual categories. Only a small set of very high-level categories is used, namely, “people”, “man-made objects”, “vegetation”, “water/sea/ocean”, “rock”, and “sky”. Some of the visual categories map to attributes, such as “outdoors”, which constrain the search. Also, the audio transcript is used to extract a topic, eg. “oil spill”. Then, a version of WordNet with entries manually annotated with visual attributes is searched for concepts supported by both image evidence and topic. For example, the detection of visual categories “man-made” and “sky” initiates a search of all subordinates of the WordNet entry “artifact” that have attribute “outdoors”, and whose definitions are compatible with the current topic “oil spill”. The evaluation of the algorithm on a single news story showed that the correct concept was in the top 20 concepts for about half of the clips.

6.4 Situated Language Learning

The idea of disambiguating which object the user is referring to using speech and image recognition is not a new idea. Deb Roy and co-authors have published numerous articles related to this idea and implemented several robotic systems. However, their focus has been on language learning and understanding, whereas our interest is in improving visual recognition. In [14], they describe a visually-grounded spoken language understanding system, an embodied robot situated on top of a table with several solid-colored objects placed in front of it on a green tablecloth. The robot learns by pointing to one of the objects, 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 things like the meaning of “blue”, “above”, “square”. The key difference between this proposal and [14] is that we want the system to recognize arbitrary objects on arbitrary backgrounds, using prior visual models of object categories as well as the user’s spoken description.

7 Timeline

Below is an approximate timeline for completing different parts of the proposed research.

February-March 2007: simulated disambiguation experiments, data collection (*completed*)

April-May 2007: collect more realistic data

June-August 2007: develop and test the disambiguation and adaptation algorithms

September-December 2007: develop out-of-vocabulary recognition algorithm

January-May 2008: integrate the final system, more testing

May 2008-August 2008: write the thesis document

References

[1] C. Chang and C. Lin: LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[2] A. Frome, Y. Singer, J. Malik, “Image Retrieval and Recognition Using Local Distance Functions”, Proceedings of Neural Information Processing Systems (NIPS) 2006.

[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] K. Grauman and T. Darrell: Approximate Correspondences in High Dimensions. In Proceedings of Advances in Neural Information Processing Systems (NIPS). 2006.

[5] K. Grauman and T. Darrell: Pyramid Match Kernels: Discriminative Classification with Sets of Image Features. MIT Technical Report MIT-CSAIL-TR-2006-020, 2006. To appear in the Journal of Machine Learning. 2006.

[6] Google Image Search, http://images.google.com

[7] A. Hoogs, J. Rittscher, G. Stein and J. Schmiederer, "Video Content Annotation Using Visual Analysis and a Large Semantic Knowledgebase," In Proceedings of CVPR, 2003.

[8] 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.

[9] S. Li and B. Wrede, "Why and how to model multi-modal interaction for a mobile robot companion," In Proc. AAAI Spring Symposium on Interaction Challenges for Intelligent Assistants, 2007.

[10] K. Murphy, A. Torralba, D. Eaton, W. T. Freeman. "Object detection and localization using local and global features", Lecture Notes in Computer Science (unrefeered). Sicily workshop on object recognition, 2005.

[11] Peekaboom Game, http://www.peekaboom.org

[12] M. Pollack et al., "Pearl: Mobile Robotic Assistant for the Elderly," In Proc. AAAI Workshop on Automation as Eldercare, 2002.

[13] 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.

[14] Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster, "A Trainable Spoken Language Understanding System for Visual Object Selection," In Proceedings of the International Conference of Spoken Language Processing, 2002.

[15] Deb Roy, Rupal Patel, Philip DeCamp, Rony Kubat, Michael Fleischman, Brandon Roy, Nikolaos Mavridis, Stefanie Tellex, Alexia Salata, Jethran Guiness, Michael Levit, Peter Gorniak, “The Human Speechome Project,” In Proceedings of the Twenty-eighth Annual Meeting of the Cognitive Science Society, 2006.

[16] B. Russell, A. Torralba, K. Murphy, and W. T. Freeman: LabelMe: a database and web-based tool for image annotation. MIT AI LAB MEMO AIM-2005-025. 2005.

[17] H. Zhang, A. Berg, M. Maire, J. Malik, “SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition”, In proceedings of CVPR, 2006.

[18] Zinger, Millet, et al, “Extracting an Ontology of Portrayable Objects from WordNet.”