Object Category Recognition Using Probabilistic Fusion of Speech and Image Classifiers

Kate Saenko and Trevor Darrell

Computer Science and Artificial Intelligence Laboratory
Massachusetts Institute of Technology
32 Vassar Street, Cambridge, MA 02139, USA
saenko,trevor@csail.mit.edu

Abstract. Multimodal scene understanding is an integral part of human-robot interaction (HRI) in situated environments. Especially useful is category-level recognition, where the the system can recognize classes of objects of scenes rather than specific instances (e.g., any chair vs. this particular chair.) Humans use multiple modalities to understand which object category is being referred to, simultaneously interpreting gesture, speech and visual appearance, and using one modality to disambiguate the information contained in the others. In this paper, we address the problem of fusing visual and acoustic information to predict object categories, when an image of the object and speech input from the user is available to the HRI system. Using probabilistic decision fusion, we show improved classification rates on a dataset containing a wide variety of object categories, compared to using either modality alone.

Keywords: multimodal fusion, object recognition, human-computer interaction.

1 Introduction

Multimodal recognition of object categories in situated environments is useful for robotic systems and other applications. Information about object identity can be conveyed in both speech and image. For example, if the user takes a picture of a cylindrical object and says: "This is my pen," a machine should be able to recognize the object as belonging to the class "pen", and not "pan", even if the acoustic signal was too ambiguous to make that distinction. Conventional approaches to object recognition rely either on visual input or on speech input alone, and therefore can be brittle in noisy conditions. Humans use multiple modalities for robust scene understanding, and artificial systems should be able to do the same.

The conventional approach to *image*-based category recognition is to train a classifier for each category offline, using labeled images. Note that *category*-level recognition allows the system to recognize a class of objects, not just single instances. To date, automatic image-based category recognition performance has

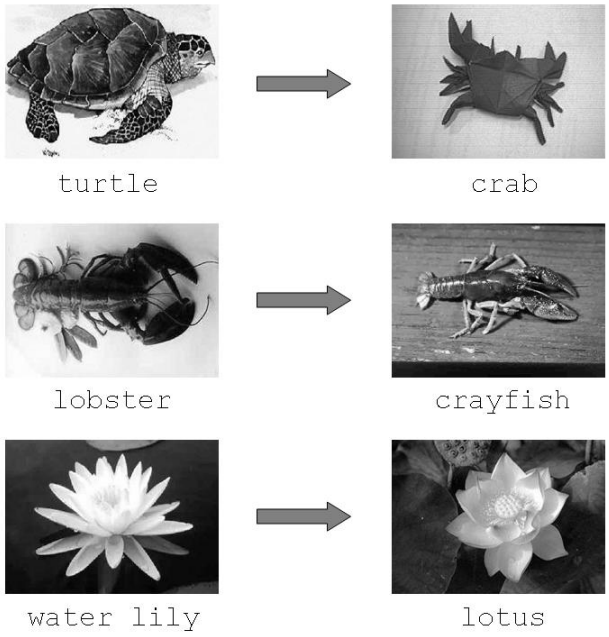

Fig. 1. Examples of the most visually confusable categories in our dataset (see Section 4 for a description of the experiments). The image-based classifier most often misclassified the category on the left as the category on the right.

only reached a fraction of human capability, especially in terms of the variety of recognized categories, partly due to lack of labeled data. Accurate and efficient off-the-shelf recognizers are only available for a handful of objects, such as faces and cars. In an assistant robot scenario, the user would have to collect and manually annotate a database of sample images to enable a robot to accurately recognize the objects in the home.

A *speech*-only approach to multimodal object recognition relies on speech recognition results to interpret the categories being referred to by the user. This approach can be used, for example, to have the user "train" a robot by providing it with speech-labeled images of objects. Such a system is described in [9], where a user can point at objects and describe them using natural dialogue, enabling the system to automatically extract sample images of specific objects and to bind them to recognized words. However, this system uses speech-only object category recognition, i.e. it uses the output of a speech recognizer to determine object-referring words, and then maps them directly to object categories. It does not use any prior knowledge of object category appearance. Thus, if the spoken description is misrecognized, there is no way to recover, and an incorrect object label may be assigned to the input image (e.g., "pan" instead of "pen".) Also, the robot can only model object *instances* that the user has pointed out. This

places a burden on the user to show the robot every possible object, since it cannot generalize to unseen objects of the same category.

We propose a new approach, which combines speech and visual object category recognition. Rather than rely completely on one modality, which can be error-prone, we propose to use both speech- and image-based classifiers to help determine the category of the object. The intuition behind this approach is that, when the categories are acoustically ambiguous due to noise, or highly confusable (e.g., "budda" and "gouda"), their visual characteristics may be distinct enough to allow an image-based classifier to correct the speech recognition errors. Even if the visual classifier is not accurate enough to choose the correct category from the set of all possible categories, it may be good enough to choose between a few *acoustically* similar categories. 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.

There are many cases in the human-computer interaction literature where multimodal fusion helps recognition (e.g. [12], [10]). 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 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 fusion 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, a limitation due to the number of categories in the labeled image database. In the future, we will explore extensions to allow arbitrary vocabularies and numbers of object categories.

2 Related Work

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. [10] use mutual disambiguation of gesture and speech modalities to interpret which object the user is referring to in an immersive virtual environment. Our proposed 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. [9] 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 recogni-

tion, 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. Note that the object recognition component is very simple, as this work focuses more on a human-robot interaction (HRI) model for object learning than on object recognition.

The idea of disambiguating which object the user is referring to using speech and image recognition is not new. In [13], the authors describe a visually-grounded spoken language understanding system, an embodied robot 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 [13] is that we focus on a realistic object categorization task, and on disambiguating among many arbitrary categories using prior visual models.

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

Murphy, et. al. present a context-sensitive object presence detection method [11]. 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 disambiguation method is somewhat similar to this, except that, in our case, the two experts operate on speech and global image features.

The current best-performing object classification methods on *Caltech 101* [3], the image database we use in our experiments, are based on discriminative multi-class classifiers. In [5], 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 [15] use a multi-class support vector machine (SVM) classifier with local interest point descriptors as visual features. We use the method of [6], which is also based on a multi-class SVM, but in combination with a kernel that computes distances between pyramids of visual feature histograms.

There has been some recent interest in using weakly supervised cross-modal learning for object recognition. For example, Fergus et al. [4] learn object categories from images obtained using a Google search for keywords describing the categories of interest. While, in this paper, we describe a supervised approach, we are also interested in exploring the idea of learning visual category classifiers in an unsupervised fashion, perhaps using web-based image search for keywords corresponding to the top speech hypotheses. This would allow an arbitrary vocabulary of object-referring words to be used, without requiring that a labeled image database exists for each word in the vocabulary.

3 Speech and Image-Based Category Recognition

In this section, we describe an algorithm for speech and image-based recognition of object categories. 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, where $k = 1, ..., C$.

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 this paper, we assume that the user always uses the same name for an object category (e.g., "car" and not "automobile".) Future work 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. In this work, 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, $f_2(x_2)$. In general, 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.

Note that 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 [2].

4 Experiments

If there is complementary information in the visual and spoken modalities, then using both for recognition should achieve better accuracy than using either one in isolation. The goal of the following experiments is to use real images, as well as recordings of users describing the objects depicted in those images, to see if such complementarity exists. Since we are not aware of any publicly available databases that contain paired images and spoken descriptions, we augmented a subset of an image-only database with speech by asking subjects to view each image and to speak the name of the object category it belongs to. Using this data, we evaluate our probabilistic fusion model. We investigate whether weighting the modalities is advantageous, and compare the mean and product combination rules.

Image Dataset. Most publicly available image databases suitable for category-level recognition contain very few object categories. The exceptions include the *PASCAL*, *LabelMe*, *Caltech101*, *ESP* and *Peekaboom* databases, which are described in [14]. 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 50 images per category. Although the categories are challenging for current object recognition methods, the task is made somewhat easier by the fact that most images have little or no background clutter, and the objects tend to be centered in the image and tend to be in a stereotypical pose. Sample images from each of the 101 categories are shown in Figure 4.

Speech Collection. We augmented a subset of the images with spoken utterances recorded in our lab, to produce a test set of image-utterance pairs on which to evaluate the fusion method. We chose the set of names W based on the names provided with the image database, changing a few of the names to more common words. For example, instead of "gerenuk", we used the word "gazelle", and so on. The exact set of names W is shown in Figure 4. 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 with 2 images from each category in the image test set, and asked to say the exact object name for each image, resulting in 12 utterances for each category, for a total of 1212 image-utterance

pairs. The reason that the images were shown, as opposed to just prompting the subject with the category name, is that some names are homonyms (e.g., here "bass" refers to the fish, not the musical instrument), and also to make the experience more natural. 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. The speech recognizer, although it was trained on an unrelated phone-quality audio corpus, achieved a word error rate (WER) of around 10% when tested on the collected category utterances. In realistic human-computer interaction scenarios, the environment can be noisy, interfering with speech recognition. Also, the category names of everyday objects are shorter, more common words (e.g. "pen" or "pan", rather than "trilobite" or "mandolin"), and the their vocabulary is much larger, resulting in a lot more acoustic confusion. Our preliminary experiments with large-vocabulary recognition of everyday object names, using a 25K-phrase vocabulary, produced WERs closer to 50%. Thus, 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. For the last two SNRs, the audio-only WERs are in a more realistic range of around 30-60%.

Training of Classifiers. We trained the image-based classifier on a standard *Caltech101* training set, consisting of the first 15 images from each category, which are different from the test images mentioned above. The classification method is described in detail in [6], here we only give a brief overview. First, a set of feature vectors is extracted from the image at each point on a regular 8-by-8 grid. A gradient direction histogram is computed around each grid point, resulting in a 128-dimensional SIFT descriptor. The size of the descriptor is reduced to 10 dimensions using principal component analysis, and the x,y position of the point is also added, resulting in a 12-dimensional vector. Vector quantization is then performed on the feature space [7], and each feature vector (block) of the image is assigned to a visual "word". Each image is represented in terms of a bag (histogram) of words. Two images can then be matched using a special kernel (the pyramid match kernel) over the space of histograms of visual words. Classification is performed with a multi-class support vector machine (SVM) using the pyramid match kernel. Our implementation uses a one-vs-rest multi-class SVM formulation, with a total of C binary SVMs, each of which outputs the visual posterior probabilities $p(c = k|x_1)$ of the class given the test image.

The speech classifier is based on the Nuance speech recognizer, a commercial, state-of-the-art, large-vocabulary speech recognizer. The recognizer has pretrained acoustic models, and is compiled using a grammar, which we set to be the set of object names W, thus creating an isolated phrase recognizer with a vocabulary of 101 phrases. This recognizer then acts as the speech-based classifier in our framework. The recognizer returns an N-best list, i.e. a list of N most likely phrase hypotheses $k = k_1, ..., k_N$, sorted by their confidence score. We use normalized confidence scores as an estimate of the posterior probability

$p(c = k|x_2)$ in Equations 1, 2. For values of k not in the N-best list, the probability was set to 0. The size of the N-best was set to 101, however, due to pruning, most lists were much shorter. The accuracy is measured as the percentage of utterances assigned the correct category label.

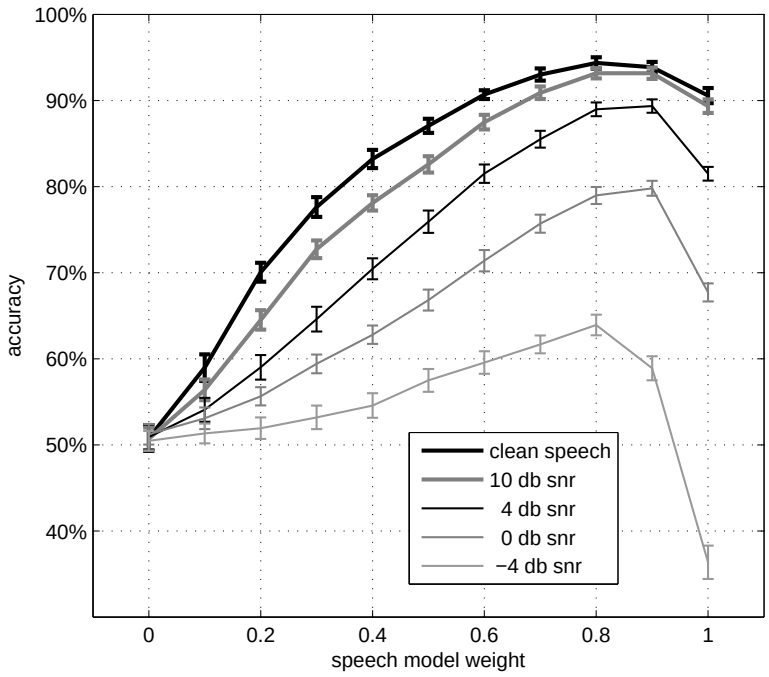

Fig. 2. Object classification using the mean rule, on the development set. Each line represents the performance on a different level of acoustic noise. The y-axis shows the percent of the samples classified correctly, the x-axis plots the speech weight used for the combined classifier.

Development and Test Sets. The test set of image-utterance pairs was further split randomly into a development set and test set. The development set was used to optimize the speech weight. All experiments were done by averaging the performance over 20 trials, each of which consisted of randomly choosing half of the data as the development set, optimizing the weight on it, and then computing the performance with that weight on the rest of the data.

Results. First, we report the single-modality results. The average accuracy obtained by the image-based classifier, measured as the percentage of correctly labeled images, was 50.7%. Chance performance on this task is around 1%. Note that it is possible to achieve better performance (58%) by using 30 training

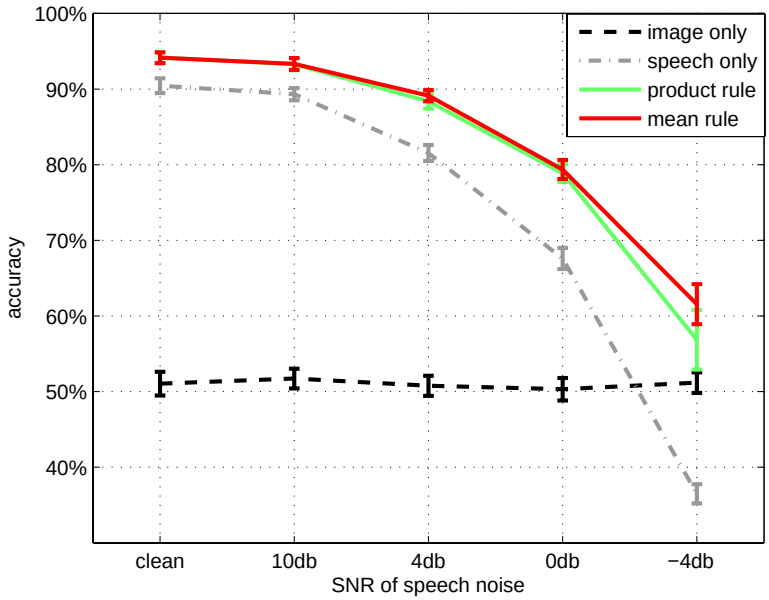

Fig. 3. Absolute improvement across noise conditions on the test set. The Y-axis shows the percent of the test samples classified correctly, the X-axis shows the SNR of the noise condition. Chance performance is around 1%.

images per category [8], however, that would not leave enough test images for some of the categories. The average 1-best accuracy obtained by the speech classifier in the clean audio condition was 91.5%. The oracle 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%.

Next, we see how the fused model performs on different noise levels. Figure 2 shows the results of the fusion algorithm on the development set, using the mean combination rule. The plot for the product rule, not shown here, is similar. Each line represents a different level of acoustic noise, with the top line being clean speech, and the bottom line being the noisiest speech with -4db SNR. The x-axis plots the speech model weight λ_2 in increments of 0.1, 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-only 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 mean rule. We also see that the weighted combination rule is better than not having weights (i.e. setting each weight to 0.5). The average accuracy on the test set, using the weight chosen on the development set for each noise condition, is plotted

in Figure 3. 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.

5 Conclusion and Future Work

We presented a multimodal object category classifier that combines image-only and speech-only hypotheses in a probabilistic way. The recognizer uses both the name of the object and its appearance to disambiguate what object category the user is referring to. We evaluated our algorithm on a standard image database of 101 object categories, augmented with recorded speech data of subjects saying the name of the objects in the images. We have simulated increasingly difficult speech recognition tasks by adding different levels of noise to the original speech data. Our results show that combining the modalities improves recognition across all noise levels, indicating that there is complementary information provided by the two classifiers. To avoid catastrophic fusion, we have proposed to use the weighted version of the mean rule to combine the posterior probabilities, and showed experimentally that there exists a single weight that works for a variety of audio noise conditions. We have thus shown that it may be advantageous for HRI 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 a proof of concept for a larger system, the first step towards multimodal object category recognition in HRI systems. We plan to continue this line of research, extending the model to handle multiple words per category, and, eventually, to extract possible object-referring words from natural dialogue. A simple extension to handle multiple object names is to map each category to a list of synonyms using a dictionary or an ontology such as WordNet. We are also interested in enabling the use of arbitrary vocabularies by learning visual category classifiers in an unsupervised fashion, using methods similar to [4]. With this approach, web-based image search would be conducted for keywords corresponding to words in the N-best list output by the speech recognizer. The returned images could then be used to build visual models for disambiguation of arbitrary objects.

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. C. Chang and C. Lin: LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
3. L. Fei-Fei, R. Fergus and P. Perona: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. IEEE. CVPR, Workshop on Generative-Model Based Vision. 2004.

4. 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.
5. A. Frome, Y. Singer, and J. Malik: Image Retrieval and Recognition Using Local Distance Functions, Proceedings of Neural Information Processing Systems (NIPS). 2006.
6. 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/
7. K. Grauman and T. Darrell: Approximate Correspondences in High Dimensions. In Proceedings of Advances in Neural Information Processing Systems (NIPS). 2006.
8. 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.
9. A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer: A multi-modal object attention system for a mobile robot, Intelligent Robots and Systems. 2005.
10. 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.
11. 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.
12. 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.
13. 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.
14. 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.
15. H. Zhang, A. Berg, M. Maire, J. Malik: SVM-KNN: Discriminative Nearest Neighbor Classification for Visual Category Recognition. In proceedings of CVPR. 2006.

Fig. 4. Sample images from the *Caltech101* database. The category name used in our experiments is shown at the top of each image.