Image Sense Disambiguation: A Multimodal Approach

by

Kate Saenko

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2009

Author .....
Department of Electrical Engineering and Computer Science
September, 2009

Certified by .....
Trevor Darrell
Associate Professor
Thesis Supervisor

Accepted by .....
Terry P. Orlando
Chairman, Department Committee on Graduate Students

Image Sense Disambiguation: A Multimodal Approach

by

Kate Saenko

Submitted to the Department of Electrical Engineering and Computer Science
on September, 2009, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

If a picture is worth a thousand words, can a thousand words be worth a training image? Most successful object recognition algorithms require manually annotated images of objects to be collected for training. The amount of human effort required to collect training data has limited most approaches to the several hundred object categories available in the labeled datasets. While human-annotated image data is scarce, additional sources of information can be used as weak labels, reducing the need for human supervision. In this thesis, we use three types of information to learn models of object categories: speech, text and dictionaries. We demonstrate that our use of non-traditional information sources facilitates automatic acquisition of visual object models for arbitrary words without requiring any labeled image examples.

Spoken object references occur in many scenarios: interaction with an assistant robot, voice-tagging of photos, etc. Existing reference resolution methods are unimodal, relying either only on image features, or only on speech recognition. We propose a method that uses both the image of the object and the speech segment referring to it to disambiguate the underlying object label. We show that even noisy speech input helps visual recognition, and vice versa. We also explore two sources of linguistic sense information: the words surrounding images on web pages, and dictionary entries for nouns that refer to objects. Keywords that index images on the web have been used as weak object labels, but these tend to produce noisy datasets with many unrelated images. We use unlabeled text, dictionary definitions, and semantic relations between concepts to learn a refined model of image sense. Our model can work with as little supervision as a single English word. We apply this model to a dataset of web images indexed by polysemous keywords, and show that it improves both retrieval of specific senses, and the resulting object classifiers.

Thesis Supervisor: Trevor Darrell
Title: Associate Professor

Acknowledgments

First of all, I would like to thank MIT and the Computer Science and Artificial Intelligence Lab (CSAIL). This dissertation would not have been possible without the support and encouragement of many people. I owe my deepest gratitude to my advisor, Professor Trevor Darrell, who inspired my interest in computer vision, and continued to encourage and motivate me throughout the years, co-advising my Master's thesis and then taking me on as his PhD student. Professor Darrell always made time for me, spending hours brainstorming ideas, and also advising me about the non-research aspects of a PhD career. Also, my heartfelt thanks go to my Master's thesis advisor, Professor James Glass, whose guidance was crucial to my early research attempts and who remained a wonderful mentor throughout my years at MIT. I am also grateful to Professor Glass and to Professor Bill Freeman for providing invaluable feedback as members of my PhD thesis committee.

I would like to thank my co-authors on this and other projects: Mario Christoudias, Timothy J. Hazen, Chia-Hao La, Karen Livescu, Louis-Philippe Morency, Michael Siracusa and Kevin Wilson. Also, my sincere thanks to my office mates: Deepti Bhatnagar, Neal Checka, Mario Christoudias, Louis-Phillipe Morency, Michael Siracusa, Sybor Wang, Kevin Wilson, and Tom Yeh. Working in the office would not have been as much fun without them!

These three amazing women were my role models while at MIT: Karen Livescu, who taught me a lot about doing research and writing papers; Ozlem Uzuner, who was a terrific friend and a dedicated audience during the preparation of this thesis, and Raquel Urtasun, who shared her keen intelligence and warm personality with everyone around her. Thank you to the members of the vision group, who provided me with valuable feedback and a positive and stimulating environment to work in: Ariadna Quattoni, David Demirjian, Gregory Shakhnarovich, Leonid Taycher, Kristen Grauman, Ali Rahimi, Jon Lee, and Andreas Geiger. Also thanks to: Maysoon Hamdiyyah, Biswajit Bose, Ghinwa Choueiter, Harold Fox, and Olya Veselova.

Finally, I would like to thank my family for their unconditional love and support

throughout this journey. To my parents: thank you for believing in me, for giving me opportunities, and for teaching me to love computer science. My mother, Tatiana, is and will always be my hero. To my dear husband René and my two lovely daughters, Isabelle and Veronika: thank you for all the love, unwavering support, patience, smiles and hugs. You are my constant inspiration.

Contents

List of Figures

List of Tables

1

Introduction

The difference between the right word and the almost right word is the difference between lightning and a lightning bug.

Mark Twain

In our daily lives, we use multiple senses to disambiguate concepts. When a phrase has several interpretations, we look at the accompanying picture. When a diagram is ambiguous, we read the caption. When we cannot see someone's face, we recognize them by their voice. In short, multimodal context is an essential part of how we learn and communicate about our world. This dissertation explores the interplay of image, speech and written language to improve the way computers learn about visual concepts.

To be useful, computer vision systems must be able to recognize a large vari-

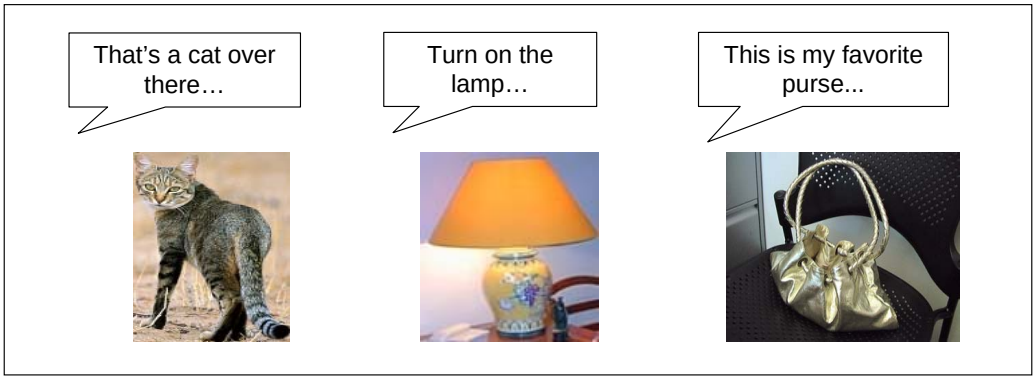


Figure 1-1: Multiple modalities can help disambiguate the identity of objects.

ety of objects. This is especially true of systems targeted towards human-computer interaction (HCI) in situated environments. Imagine a user communicating with a household robot about objects in her home, as illustrated in Figure 1-1. The user would expect such a robot to understand what objects she is referring to. However, this is problematic from the computer vision perspective. It is estimated that humans can recognize tens of thousands of object categories [Biederman, 1987]. In computer vision research, attempts to construct object recognizers for more than a few hundred categories have been stymied by the lack of labeled image data that is required to train most state-of-the-art methods [Liu *et al.*, 2007]. In HCI research, the goal of large-vocabulary automatic speech recognition has been achieved, but robustness to background noise remains an issue [Potamianos *et al.*, 2003].

The Internet is an enormous source of free data. To avoid manually labeling training examples for each object category, computer vision researchers have turned to web search engines as a cheap source of training images. However, while thousands of images are available at the click of a mouse button, their precision is low due to the ambiguity of text queries [Fergus *et al.*, 2005]. Figure 1-2 illustrates this by displaying the mixed-sense images returned by the Yahoo!^{TM} image search engine for the query "mouse". The images depict a medley of visual senses, from computer mice to field mice to Mickey Mouse. Such mixed results are due in part to *polysemy*, i.e. words having multiple meanings, and also to the inherent imprecision of keyword-based image indexing performed by search engines. Thus, an attempt to overcome

the problem of data scarcity has resulted in a new dillemma – that of image sense disambiguation.

To address these challenges, this dissertation contributes two main solutions. The first solution uses multimodality in new ways to reduce the need for supervision and to increase the robustness of object recognition. The second solution takes Internet-based image harvesting methods to the next level, tapping into the vast pool of human-generated online knowledge to automatically disambiguate image senses.

Combining multiple modalities, views, or feature streams to achieve greater robustness is a popular paradigm in artificial intelligence (AI) research. Similarly, there has been a recent rise in the number of methods that leverage multiple views to reduce supervision [Blum and Mitchell, 1998], [Yu *et al.*, 2008]. What is different about this dissertation's multimodal approach is that it targets the problem of object recognition using modalities seldom utilized for that purpose, namely speech and written language.

Most traditional approaches to object recognition are based solely on the pixel content of the image. However, with the advent of cheap digital cameras and fast Internet connections, more and more high quality multimedia data is becoming available online. In such data, objects depicted in images are often accompanied by related audio, e.g., in instructional videos, product reviews, or educational programs. Furthermore, spoken utterances referring to objects depicted in the image or video can arise in human-computer interaction, robotics, voice-tagging, question-answering, and other applications. The existing unimodal (speech-only or pixel-only) approach to object reference resolution can be unreliable. As an example, short words such as "pen" and "pan" are easily confused acoustically. Imagine showing your robot where you store your pens, only to find it attempting to cook with one later! By combining speech and image signals, we leverage their complimentary nature and facilitate recognition.

Another way in which this dissertation uses non-traditional information sources to aid object recognition is in exploiting the words surrounding images on the web. Text co-occurring with images on a web page can be a rich cue as to which object is actually depicted in the image. For example, even before we see the image re-

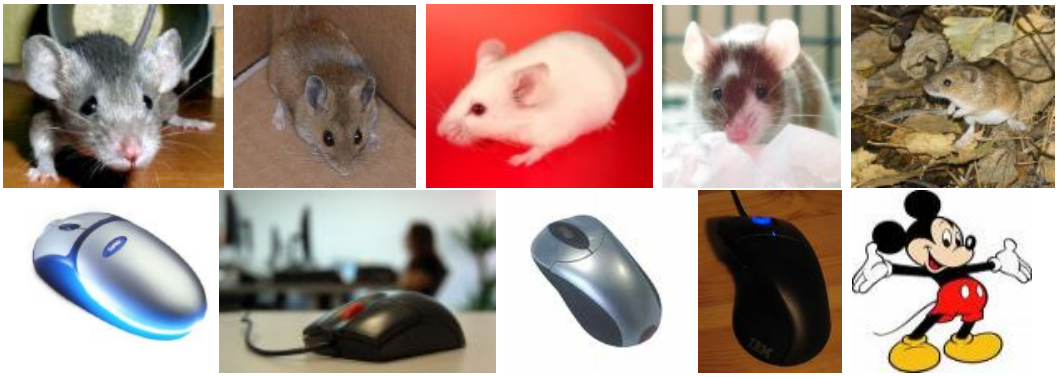


Figure 1-2: **Which sense of "mouse"?** Mixed-sense images returned from an image keyword search.

turned for the query "watch", we might know from the discussion of tornados that preceeds it on the web page that it is probably not a wristwatch photo. While several models have been proposed for images and their (manually generated) captions ([Barnard and Johnson, 2005],[Blei and Jordan, 2003], [Jain *et al.*, 2007]), our approach is different in that it learns image senses from free-form web pages. The abundance of free, unlabeled, unstructured information on the web enables us to learn models of image sense without manual annotation.

The reader may well be wondering, how is it possible to learn a model of image sense without *any* labeled examples? The answer is that we once again employ linguistic information sources, this time using word definitions extracted from online dictionaries or encyclopedias. The cornerstone of our approach is in combining these online knowledge repositories with unlabelled web text and images in order to ground models of image senses.

1.1 Thesis Contributions

The main contribution of this thesis is a multimodal framework for learning a large vocabulary of visual senses with a minimal amount of supervision. Previous methods have either required labeled examples or assumed a single sense per word. In contrast, this dissertation addresses the problem of ambiguous word meanings. We present a

succession of models, each of which represents a stepping stone toward the goal of unsupervised, on-the-fly learning of objects in interactive situations.

For the case when a small number of labeled images are available, Chapter 4 develops a semi-supervised model of image sense. This model extends previous image-only clustering methods to include text features gleaned from the context of the originating web page. It develops a novel approach to combining text and image features using latent dimensions. Experiments on a dataset collected by searching for images on the web using polysemous queries demonstrate the advantage of the combined text and image approach over single-view baselines.

For the case when no labeled images are available, Chapter 5 proposes a novel unsupervised image sense model that can learn a model of an arbitrary visual concept. The algorithm requires no labels; the only input required from the user is a dictionary entry corresponding to a visual concept. Web search data and an online dictionary are exploited to develop a generative probabilistic model of image sense. Applications of the model include web image sense disambiguation and image-based object classification. An additional contribution of this approach is the ability to learn not just an image-based object model, but also a language-based model of the corresponding word sense. The latter could be used for discourse processing and language understanding in an integrated system.

Relaxing the supervision assumptions further, Chapter 6 presents a method that automatically selects only the *visual* senses for inclusion in the object model for a particular word, and filters out the abstract senses. This sense selection is based on known semantic relations between words and enables retrieval of broad classes of visual concepts (e.g. animals, people, etc.) The applications include object classification, as well as retrieval of the visual senses of a keyword from web search results.

A final contribution, described in Chapter 7, is a method that uses both the spoken reference to an object and the image of that object to recognize its identity. The method combines a speech recognizer and an image classifier in a single framework. We demonstrate improvements over unimodal recognition on a fixed vocabulary of objects, using a dataset of images paired with spoken utterances.

Before presenting the methods, we first discuss related work areas in the next chapter, and then describe the datasets used for evaluation in Chapter 3.

2

Related Work

The ideas in this thesis draw on several areas of research, including semantic image retrieval, object recognition, word sense disambiguation, and latent topic models. This thesis focuses on a subset of methods which utilize unlabeled and weakly labeled data, such as images collected from the web. While semantic image retrieval and object recognition are two closely related subfields of image processing, a full review of these methods is beyond the scope of this work. Section 2.1 will review related work in the semantic retrieval community, as well as papers in the object recognition community dealing with harvesting image datasets from the web.

Latent topic models have been applied widely to the problem of clustering text documents, and, more recently, images. Section 2.2 will provide some background on latent topic models, starting with single-view models of bag-of-words data such as latent Dirichlet allocation, then moving on to models of captioned images, and finally

describing methods that apply topic models to web-based dataset construction. The rest of the chapter will describe work in the area of word sense disambiguation from text and images (Section 2.3), as well as the area of multimodal reference resolution, where image, speech and gestures are used to resolve object references (Section 2.4). Finally, we summarize closely related ideas and contrast them to our own work.

2.1 Image Retrieval and Object Recognition

Content-based image retrieval (CBIR) and object recognition have originated as two separate fields of research, but have evolved over the past decade to become very closely related. The main difference is in the formulation of the problem: image retrieval focuses on searching a given image collection for images matching a user's query, while object recognition focuses on finding a specific object in a given image. The goal in CBIR is to match the entire image, rather than a single object present in it; however, recognizing objects contained in the image can be an intermediate step of CBIR, and, conversely, object recognition can be formulated as scene recognition. In this section, we will give a brief overview of each area, and concentrate on methods that are most relevant to this work.

2.1.1 Semantic Image Retrieval

Answering a user's query lies at the heart of image retrieval. The main approaches are text-based and content-based retrieval. The earliest text-based approaches date back to the 1970s, and involve searching a database of manually annotated images for a matching keyword. Most modern web image search engine, such as Google or Yahoo, still use the text-based approach, searching for the query word in the image filename and webpage text. Content-based image retrieval was introduced in the 1980's and involves matching a query image based on either its low-level features, such as color, texture and shape, or the high-level concepts present in the image, such as 'sky', 'water', 'animals', etc. Matching based on high-level concepts is referred to as semantic image retrieval. The query can be either an image, or a text query such

as "find me a picture of a beach", or an even more challenging request such as "find me a picture of a parent hugging a smiling child". For an extensive review of CBIR we refer the reader to [Liu *et al.*, 2007].

While semantic retrieval borrows techniques from object and scene recognition, such techniques do not provide the complete solution to CBIR. The use of supervised object recognition to automatically label images with high-level concepts for later use in retrieval has generally been limited to color categories (e.g. 'red', 'blue'), uniformly colored and textured concepts (e.g. 'sky', 'grass', 'water'), and a small number of object categories (e.g. 'faces', 'cars'). One problem with supervised object recognition is that it requires many training images for every concept, and currently has only been shown to work reliably for fewer than a few hundred objects. To deal with this limitation, semi-supervised methods can be used for retrieval. For example, in [Feng *et al.*, 2004], co-training is used to iteratively bootstrap two separate classifiers trained on a small number of labeled images and their text contexts. The paper reports a level of performance similar to the fully supervised version of the approach, but using much fewer labeled examples.

Unsupervised learning can also be applied to cluster the images for use in retrieval. The hypothesis is that semantically similar images will cluster together. However, traditional methods such as k-means often fail to cluster images into different concepts [Liu *et al.*, 2007], although this must depend greatly on the similarity measure used. The CLUE retrieval system proposed in [James *et al.*, 2003] uses a more sophisticated clustering method called NCut to retrieve clusters of images similar to the query.

Another common approach is to improve retrieval accuracy by introducing a user feedback loop into the process. Relevance feedback (RF) typically work by adapting the similarity measure to better retrieve the images that the user has indicated as being relevant to the previous queries. The advantage of RF is that it can learn the user's intentions on the fly, by providing a "more like this" option.

CBIR can been applied to diverse image collections including personal photos, specialized image databases such as medical or art galleries, and annotated datasets such as COREL [Corel, 2009]. However by far the largest repository of images is the

World Wide Web. Currently, most commercial search engines only retrieve images based on the text query, however, a few steps in the direction of more content-based retrieval have already been taken. For example, Google Similar Images [Google, 2009] allows the user to re-organize the retrieved images by content (currently limited to 'news', 'faces', 'clip art', 'line drawings' and 'photos'), by color, or by similarity to one of the images. Still, web image search for semantic concepts has low precision, and typically returns several topics mixed together [Cai *et al.*, 2004].

Recently, there has been an interest in using both the image features and the HTML content of the returned results to alleviate the poor precision of web search. In CBIR literature, several methods have been proposed to re-organize the results returned by a search engine to improve query by keyword [Feng *et al.*, 2004] and query by example [James *et al.*, 2003], or to make it easier for the user to find desired images [Cai *et al.*, 2004]. Meanwhile, in the object recognition community, efforts to avoid manually labeling training examples of each object category have led researchers to use the search engines as a cheap source of training data. We will describe these approaches in Section 2.1.3, but first, let us briefly survey the standard labeled datasets used in the object recognition field.

2.1.2 Datasets for Object Recognition

At the time of this writing, several labeled image datasets are available to researchers for the development of object recognition algorithms. However, none satisfy the requirement of providing images of *any* object named by the user. Furthermore, by nature they are *static* requiring more manual effort to be expended should a new object category become necessary.

Most hand-collected datasets contain only a few categories [Ponce *et al.*, 2006]. Larger datasets freely available to the object recognition community contain several hundred categories: Caltech-101 [Fei-Fei *et al.*, 2007], Caltech-256 [Griffin *et al.*, 2007], PASCAL [Everingham *et al.*,], LotusHill [Yao *et al.*, 2007], and
[Fink and Ullman, 2008]. The two exceptions, containing several thousand categories each, are LabelMe [Russell *et al.*, 2008] and ImageNet [Deng *et al.*, 2009]. Finally,

the Tiny Images project [Torralba *et al.*, 2008] has collected 80 million thumbnail-sized images for about 75,000 keywords by crawling the Internet. While its size is impressive, it is not suitable for evaluation because it does not contain labels.

LabelMe is a dataset collaboratively collected through an open web-based annotation tool [Russell *et al.*, 2008]. It contains natural images of indoor and outdoor scenes, many of which are annotated by users of the dataset. An annotation consists of the outline of an object appearing in the image and a free-text description (e.g. "car") Russell et al. [Russell *et al.*, 2008] reports that, as of 2006, it contained more than 4,000 unique descriptions and that while the rate of new outlines being added was increasing, the rate of new descriptions has been slowing down. They also report that most descriptions had fewer than 10^3 instances. We conducted an informal experiment to assess the coverage of common household objects by searching the online interface. A query for "pliers" returned 1 image, for "cellphone" 11, for "telephone" 349, for camera 67, and for "mug" 362. Many of the objects were of rather low resolution. These results reflect the fact that the database is designed to provide annotations of *whole scenes* and objects in the context of natural images. However, the low number of instances and poor resolution of many categories makes it difficult to use LabelMe for object recognition out of context.

ImageNet. In addition to covering a large vocabulary of words, another goal of this work is to build sense-disambiguated models. Very few existing datasets contain sense-disambiguated labels. A notable exception is the ImageNet [Deng *et al.*, 2009] project. ImageNet is a new project that aims to collect labeled images of the 80,000 synsets of WordNet, with an average of 500-1000 for each synset. The images are collected by searching the web for the synset words and paying human labelers to select good examples from among the returned results. It is so far the most ambitious dataset collection effort, expected to result in tens of millions of images. The fact that it contains sense-disambiguated labels makes it a good dataset for evaluating our unsupervised method for collecting sense-specific data, although since it does not include webpage text, we can only use it to test image-only classification. Unfortunately ImageNet was released after the experimental portion of this dissertation was

completed, and testing on it remains for future work.

2.1.3 Image Harvesting for Object Recognition

While object recognition has been shown to work well for several visual categories, such as human faces, the lack of labeled datasets containing many examples per class and large intra-class variation has thus far prevented it to be successful at recognizing a wide variety of categories [Liu *et al.*, 2007]. Efforts to avoid manual labeling by typing the name of the object into a search engine and automatically gaining access to thousands of training images were met with low precision due to the ambiguity of word meanings [Fergus *et al.*, 2005]. Several approaches to dealing with the precision problem have been proposed, including iterative re-ranking with a classifier trained on the top-ranked images [Schroff *et al.*, 2007], bootstrapping image classifiers from labeled image data [Li *et al.*, 2007], clustering the returned images into coherent components [Fergus *et al.*, 2005],[Li *et al.*, 2007],[Berg and Forsyth, 2006], and incorporating user feedback [Collins *et al.*, 2008],[Berg and Forsyth, 2006].

Several web-based dataset construction methods have incorporated both the image and the HTML features of search results [Schroff *et al.*, 2007],[Berg and Forsyth, 2006]. Schroff et al. [Schroff *et al.*, 2007] first used a Bayes classifier based only on text and HTML metadata (such as whether the keyword appears in the URL, the ALT tag, etc.) to re-rank the images returned from web search. The classifier was trained in a category-independent manner to predict whether the desired object appears in the image. The images ranked highest by the text classifier were then used as noisy training data for a support vector machine, which was used in turn to re-rank the set once again to improve semantic retrieval, and to classify unseen examples. The evaluation showed that the combination of text and metadata with image features improves re-ranking over using either view alone.

A relevance-feedback approach was proposed by Collins et al. [Collins *et al.*, 2008], who asked the user to label several dozen images chosen randomly from search results, and iteratively re-trained their boosting classifier on those images.

Recently, the object recognition field as a whole has seen a trend towards repre-

senting images as bags of visual words. The parallels between the problem of clustering such bag-of-words data and problems in the text processing community have led to the application of traditionally text-only topic models to unsupervised object recognition. In fact, most of the existing web-based dataset construction methods incorporate topic models. In the next section, we first provide the reader with a review of topic models as they were introduced in the text processing literature, and then discuss their extensions to models of images and the text associated with them.

2.2 Latent Topic Models

Latent topic analysis is a classic problem in natural language processing (NLP) and information retrieval. In the 1980s, latent semantic analysis (LSA), sometimes also referred to as latent semantic indexing (LSI), was developed [Deerwester *et al.*, 1990]. LSA applies singular value decomposition to a term-document matrix to find latent "concepts" in the document corpus. Although LSA is a widely used model in NLP, generative latent variable models have recently been suggested as a more principled approach to probabilistic modeling of documents. Such approaches include mixtures of unigrams [Nigam *et al.*, 2000], probabilistic LSA [Hofmann, 1999], and latent Dirichlet allocation [Blei *et al.*, 2003]. The advantages of using a proper generative model is that standard techniques can be used for inference, parameter estimation, and model combination.

2.2.1 Probabilistic Latent Semantic Indexing

Probabilistic LSI (pLSI), introduced by Hofmann [Hofmann, 1999], discovers hidden topics, or distributions over discrete observations (such as words), in unlabeled data. It models a document collection using a graphical model with documents and words represented as observed variables and "topics" as hidden variables. The topics represent distributions over word counts. In practice, they tend to align with coherent themes within the corpus and can help to automatically distinguish between different meanings and uses of the same word [Hofmann, 1999], a key reason for their choice

to perform visual sense disambiguation in this thesis.

The pLSI model makes the conditional independence assumption that, given the latent topic z, the word w is generated independent of the identity of the document d in which it occurs. In contrast to mixture models, the document is not assigned to a cluster, but rather represented as a list of topic proportions. Given a collection of N documents, each containing a bag of N_d words with vocabulary size M, pLSI assumes the following generative process:

1. pick a document $d \in 1, ..., N$ with prior probability $P(d)$,

2. for each word token i, sample a latent topic $z_i \in 1, ..., K$ from $P(z|d)$, a multinomial distribution with parameter θ_d,

3. choose a word $w_i \in 1, ..., M$ from $P(w|z)$, a multinomial with parameter ϕ^{z_i}.

The probability of generating a word in a document is

$$P(d, w|\phi_{1:K}, \theta_d) = P(d)\ P(w|d, \phi_{1:K}, \theta_d) = P(d) \sum_{z=1}^{K} P(w|z, \phi_{1:K})\ P(z|\theta_d) \qquad (2.1)$$

The probability of generating a document consisting of words $w_1, ..., w_{N_d}$ is

$$P(d, w_1, ..., w_{N_d}|\phi_{1:K}, \theta_d) = P(d) \prod_{i=1}^{N_d} \sum_{z=1}^{K} P(w_i|z, \phi_{1:K})\ P(z|\theta_d) \qquad (2.2)$$

The variable d represents an index into the list of training documents. The parameters of the pLSI model are the document-specific topic distributions $P(z|d)$, learned for each document d in the corpus, and the topic-specific word distributions, $P(w|z)$. EM can be used to estimate the parameters for a collection of documents. The pLSI model suffers from two shortcomings: 1) the number of parameters grows linearly with the number of documents, and 2) there is no natural way to generate previously unseen documents. (One way is to use EM in a "fold-in" procedure, keeping $P(w|z)$ fixed, to estimate $P(z|d)$ for an unseen document.) These problems can lead to overfitting, and a solution was proposed in the form of Latent Dirichlet Allocation (LDA).

2.2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) was introduced by (Blei et al.[Blei *et al.*, 2003]). It is a fully generative model that is similar to the pLSI model, but treats the topic distribution as a hidden variable sampled from a Dirichlet prior. This prior, along with an additional prior over the conditional distribution of words given the topic, are shared across all documents in the corpus. As in pLSI, each document is modeled as a mixture of topics $z \in 1, ..., K$. A given collection of N documents, each containing a bag of N_d words, is assumed to be generated by the following process:

1. for each topic $j = 1, ..., K$, sample the parameters of a multinomial distribution over words ϕ^j from the Dirichlet prior with parameter β,

2. for each document d, sample the parameters θ_d of a multinomial distribution over topics from the Dirichlet prior with parameter α,

3. for each word token i, choose a topic z_i from the multinomial θ_d, then choose a word w_i from the multinomial ϕ^{z_i}.

The probability of generating a document is

$$P(w_1, ..., w_{N_d}|\phi, \theta_d) = \prod_{i=1}^{N_d} \sum_{z=1}^{K} P(w_i|z, \phi)\ P(z|\theta_d) \tag{2.3}$$

To perform inference, a number of approximate inference algorithms can be applied. In the experiments in this dissertation, we use the Gibbs sampling approach of [Griffiths and Steyvers, 2004] to produce samples from the posterior distribution $P(z|w)$. The θ and ϕ variables are integrated out using the conjugate Dirichlet priors. Given a posterior sample, the parameters are then estimated from their predictive distributions conditioned on the data as:

$$\hat{\phi}_w^j = \frac{n_w^j + \beta}{\sum_{v=1}^{M}(n_v^j + \beta)}\ , \qquad \hat{\theta}_d^j = \frac{n_d^j + \alpha}{\sum_{k=1}^{K}(n_d^k + \alpha)}\ , \tag{2.4}$$

where n_w^j is the number of times word w was assigned to topic j, and n_d^j is the number of words in document d assigned to topic j. Using symmetric Dirichlet

priors with scalar hyper-parameters α and β has the effect of smoothing the empirical distribution. The number of topics K is a fixed parameter that can be set by cross-validation. The hierarchical Dirichlet process (HDP) model addresses this issue by marginalizing over K [Teh *et al.*, 2003].

Learning the latent structure is only the first step in solving the problem at hand. The second step is to compute a similarity measure between pairs of documents, given the learned latent structure, for the purpose of re-ranking the retrieved documents. The basic approach to using a hidden topic model for retrieval is to score each document by the likelihood of generating the query document under its model. We follow the approach described in [Wei and Croft, 2006] to perform retrieval with the LDA model. Given a query document q consisting of words $w_1, ..., w_{N_q}$, and a document d with estimated posterior parameters θ_d and ϕ, the similarity measure is the (log) likelihood of the query under d's model:

$$D(q, d) = log(P(w_1, ..., w_{N_q}|\phi, \theta_d)) \tag{2.5}$$

which is given by ((2.4)).

LDA has been successfully applied to a variety of natural language tasks, such as analyzing, organizing and searching document collections. It has also been extended in a variety of ways. For example, the Author-Topic model [Rosen-Zvi *et al.*, 2004] extends the LDA model to include the identity of the authors of the document. The hidden Markov model (HMM) LDA [Griffiths *et al.*, 2004] separates syntactic words from content words by making the word distribution dependent on the previous word's syntactic state. For example, this allows the model to label the instance of the word "and" as a syntactic word in "observe and measure", and as a content word in "draw an AND gate". HMM-LDA has been applied by [Hsu and Glass, 2006] to the problem of adaptive language modeling in lecture transcription. In audio lecture processing, one of the difficulties is that the data available to train the language model (i.e. probability of the speaker saying a particular word) either matches the domain (lecture topic) or the style (written text vs. spoken lecture) of the test data,

(a) water grass flowers trees sky

(b) "On the way west the expedition towed, poled, and rowed their boats up the Missouri River, against the current, on a good day making 20 miles (32 kilometers)"

Figure 2-1: Unlike image captions (a), the text surrounding a web image (b) does not generally consist of words corresponding to each image region.

but rarely both. However, an effective language model needs to not only capture the sponteneous speaking style of a lecturer, but also the domain-specific vocabulary. By applying the HMM-LDA model, [Hsu and Glass, 2006] were able to dynamically adapt the language model to the apparent topic substructure of a lecture.

2.2.3 Latent Topic Models of Annotated Images

In semantic retrieval literature, latent topic models have been applied to model image and caption data (e.g. [Barnard *et al.*, 2003],[Blei and Jordan, 2003]). It is important to note that the free-text contents of HTML pages that co-occur with images retrieved by web search are different from captions (see Figure 2-1). Image captions are produced by a human labeler, who labels each image in a database (e.g. COREL [Corel, 2009]) with the purpose of describing the image content. An example caption is "rocks,sky,grass". The goal is usually to provide at least one word describing each concept appearing in the image. Therefore, models of such annotated data can assume that there is a correspondence between each image region and one of the words in the caption.

In models of captioned images, image annotation is performed by using the conditional distribution of words given an image region to predict the most likely word,

which serves as a category label [Blei and Jordan, 2003]. However, in the context of web images, it does not make sense to predict noisy free-text words that are not meant to be labels. Certain words on the page, for example, the words "Mickey Mouse" co-occurring with an image returned for the query "mouse", might serve as captions, however, such descriptive words are rare and are not necessarily related to the desired object category. In general, a text context word does not have a corresponding visual region, and vice versa. Thus, existing image and caption processing methods address a different problem than the one we are interested in. Nevertheless, it is a closely related problem, so we will briefly review those methods that are based on topic modeling.

Two extensions of LDA were introduced by Blei to model annotated data [Blei, 2004]. The first is a Gaussian-multinomial LDA (GM-LDA), which uses latent topics to represent the joint clustering of image regions and caption words. An image/caption pair is generated by first sampling θ from the Dirichlet prior, then, for each image region, choosing a topic z and sampling a region descriptor from a Gaussian distribution conditioned on z. After the image regions have been generated, each caption word is sampled by first choosing a topic v and then sampling a word from the multinomial distribution conditioned on v. While both the image topics z and the caption topics v are generated from the same distribution, there is no explicit dependency between them, so it is possible for them to form two distinct sets.

Specific correspondence between image regions and caption words is introduced by the correspondence-LDA (Corr-LDA) model. It forces the image and word topics to be associated with each other by selecting each word topic from one of the topics that generated the image regions. Specifically, it proceeds identically to GM-LDA to sample N image regions, and then, for each caption word, it chooses a region index y from a uniform distribution on the interval $(1, ..., N)$, and selects a word from the multinomial distribution conditioned on y's topic.

In work closely related to Corr-LDA, a People-LDA [Jain *et al.*, 2007] model is used to guide topic formation in news photos and captions, using a specialized face recognizer. The caption data in news photos is less constrained than annotations

and includes some non-category words, however, it is still far more constrained than free-text web pages.

2.2.4 Latent Topic Models of Image Search Results

Fergus et al. [Fergus *et al.*, 2005] were the first to attempt to train a classifier given nothing but the name of the object. They clustered images retrieved by a search engine for that name using pLSI adapted to include the position of the topic in an image. The assumption behind their method is that one of the learned topics will align closely with either the desired object. The foreground object topic was selected using a small validation dataset, which was automatically collected as follows: the word was translated automatically into several languages, and the first page of image search results from each language was used to create a high-precision set. The approach did not include text features and was limited by the assumption that a single topic would be formed by the desired object features.

Berg et al. [Berg and Forsyth, 2006] also used a probabilistic topic model (namely, LDA) to cluster the retrieved images, but based on the text words surrounding the image link. User feedback was then sought by asking which clusters belonged to the desired image category and which were unrelated. The images in the labeled clusters were then used to train a voting classifier.

Bootstrapping methods rely on the presence of good initial classifiers trained on seed labeled data. The bootstrapping approach named OPTIMOL [Li *et al.*, 2007] used the Caltech-101 object dataset to obtain initial seed data, and iteratively refined a classifier based on a hierarchical Dirichlet process.

2.3 Sense Disambiguation

There is a reason why retrieving visual concepts based only on the text name has such poor precision – words are ambiguous! While a human can read the sentence "The bank is closed" and guess that it is probably talking about a financial institution, a computer would have a much harder time disambiguating the sense of the

word "bank". Context, either textual of visual, can help automatic methods in this task. For example, a news image showing the hotel heiress Paris Hilton can help disambiguate the caption "Paris Hilton" as referring to the person, not the hotel.

2.3.1 Sense Disambiguation in Text

The problem of Word Sense Disambiguation (WSD) by computer is one of the most fundamental problems in natural language processing. The difficulty lies in the fact that disambiguation often requires common sense. Word sense disambiguation is considered an AI-complete problem, meaning that it's at least as hard as the most difficult problems in artificial intelligence [Navigli, 2009]. That said, statistical approaches that do not attempt to understand language but rather make a decision based on the statistics of the surrounding words have had some success in WSD. The current state of the art is about 75% accuracy in word sense disambiguation with supervised learning. For a complete review of traditional (natural language) WSD the reader is referred to [Agirre and Edmonds, 2006, Navigli, 2009].

One of the problems with obtaining supervised training data for WSD is the difficulty in training people to tag word senses. This has led to the advent of semi-supervised and unsupervised algorithms. Yarowsky [Yarowsky, 1995] proposed an unsupervised bootstrapping method, and suggested the use of dictionary definitions as an initial seed. The algorithm uses the "one sense per discourse" property of natural language, where a polysemous word will typically take on only one meaning in a given discourse. It accumulates a list of *collocations*, or words that co-occur with the target word more frequently than would be expected by chance, that are indicative of each sense. For example, the collocation "computer" and "mouse" is highly indicative of the "input device" sense of "mouse".

2.3.2 Sense Disambiguation in Text and Pictures

The saying "a picture is worth a thousand words" certainly seems to apply to word sense disambiguation, at least for a human observer. For computer methods, images

that co-occur with text may also be helpful, even though computers' image understanding abilities are limited. To investigate whether using both images and text would improve WSD performance, Barnard and Johnson [Barnard and Johnson, 2005] created a new dataset of Corel images [Corel, 2009] paired with text passages from a sense-disambiguated text corpus. Each image was chosen such that it illustrated the corresponding text passage, similar to a news caption and accompanying photo. A statistical model linking image regions to word senses was trained on the training portion of the image/caption data. Given a test image and caption, the model predicted the most likely word sense for each word in the caption based on the test image. The evaluation showed that this approach exceeded the performance of two text-only WSD methods, and that combining image-based and text-based methods resulted in further improvement.

While images and their captions are an interesting domain, images occurring "in the wild", i.e. on the world wide web, typically lack captions (see Section 2.2.3). We are not aware of work on WSD applied to general webpage text and images. The work of Loeff et al. [Loeff *et al.*, 2006] introduced *image sense disambiguation (ISD)* for web images retrieved by an internet search engine for an ambiguous keyword. There are several distinctions between WSD and ISD, according to [Loeff *et al.*, 2006]. First, because the keyword indexing is not done on human generated labels, web images contain not only the *core* senses, but also *related* meanings. For example, the keyword "watch" can retrieve images of watch mechanisms, watch straps, people wearing watches, etc. Furthermore, the authors make a distinction between *iconographic* senses within a single core word sense: pictures of the fish "bass" may include zoological illustrations, swimming fish, caught fish, cooked fish, etc. Finally, the image may not contain any core or related senses at all. The authors address the problem of distinguishing between core, related, and unrelated senses of web search results by performing spectral clustering in both the text and image domain. Evaluation consisted of computing how well the clusters matched human-annotated senses. No classification was done to assign labels to clusters.

Word sense disambiguation with words and pictures is not the only research prob-

lem in AI where meaning must be inferred using input from multiple modalities. In human-computer interaction, such problems come up when the interface provides different means for the user to interact with the system, such as using speech, gesture, gaze, etc. One issue is that of multimodal reference resolution, or identifying the object referred to by the user with varying degrees of information about it contained in each modality.

2.4 Multimodal Reference Resolution

Multimodal interaction using speech and gesture dates back to Bolt's Put-That-There system [Bolt, 1980]. Since that pioneering work, there have been a number of projects on virtual and augmented-reality interaction combining multiple modalities for reference resolution. In HCI, ambiguities can arise in difference forms, leading to problems of *deixis*. Deixis is a phenomenon wherein understanding of the meaning of certain words and phrases requires contextual information [Wikipedia, 2009]. For example, English pronouns (e.g. "he", "it") and place references (e.g. "this city") require resolution to concrete people, things and places. Another type of deixis occurs when the user refers to objects in the environment, for example, points to a cup and says "This cup is hot". In this case, the ambiguity is not only in *which* physical cup the user means, but also what sense of "cup" he means and what the object looks like. Since speech recognition is prone to errors, the system may have trouble recognizing the word "cup". In work by Kaiser, et al. [Kaiser *et al.*, 2003], mutual disambiguation of gesture and speech modalities to interpret which object the user is referring to in an immersive virtual environment. However, in the virtual reality and game environments, the identity of surrounding objects is known, making the problem easier. In the real-world interaction scenarios that we are interested in, object identity and location is unknown and must be determined based on visual appearance.

Haasch, et. al. [Haasch *et al.*, 2005] 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 track-

ing, 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. 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. We believe that such a system would benefit from being able to recognize more complex types of visual concepts.

The idea of disambiguating which object the user is referring to using speech and image recognition has also been studied by Roy et al. In [Roy *et al.*, 2002], 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 visual concepts, such as the meaning of "blue", "above", "square". Again, the concepts considered are simple, and the focus of the work is on language learning rather than object recognition.

2.5 Multimodal Fusion and Classifier Combination

Throughout this dissertation, the issue of combining evidence from multiple sources comes up again and again. In this section, we discuss the main approaches to multimodal fusion and classifier combination at a very coarse level, without delving into the multitude of practical combinations schemes developed for the various modalities, feature sets, and classifiers.

Multiple feature streams can be extracted from different modalities, e.g. speech and vision, or from different views of the same modality, e.g. pixel intensities and motion between video frames. The key is that the feature streams be compementary in some way. At the most general level, multimodal approaches fall into two broad categories: early fusion and late fusion. In early fusion, the features extracted from different sources are concatenated together to form a single feature vector. In contrast, late fusion approaches combine the decisions of separate classifiers, each of which is based on a distinct feature representation. For example, in audio-visual speech recognition, one early fusion method is to stack the acoustic features together with visual features extracted from moving lip images into a single feature vector (see [Potamianos *et al.*, 2003], for example). In [Saenko *et al.*, 2005, Saenko *et al.*, 2009], we have explored a late fusion approach to audio-visual speech recognition, where acoustic and visual features are observed at separate nodes in a graphical model, with other nodes performing the combination.

While multimodal problems such as audio-visual speech recognition involve synchronized feature streams which are actually different manifestations of the same underlying cause (the movement of speech articulators), other multimodal feature sets are less tightly coupled. For example, when we think about combining spoken references to objects with image features, concatenating the two feature spaces does not seem to make sense. Besides being more appropriate for decoupled features, late fusion has several additional advantages. One is efficiency: one can use the more efficient classifier to quickly eliminate most of the hypotheses, and only then apply the computationally intensive classifier [Kittler *et al.*, 1998]. For example, a speech recognizer can be used to efficiently eliminate most of the word hypotheses in a large vocabulary before using a visual classifer to disambiguate between the remaining few words (see Chapter 7. The other advantage is the greater accuracy resulting from the commonly used classifier combination rules: (weighted) sum rule, (weighted) product rule, min and max rules, and voting [Kittler *et al.*, 1998].

True to their names, the product rule multiplies the posterior probabilities of the class given by each classifier, while the sum rule adds them. The contribution of

each modality/classifier can also be weighted to reflect its reliability, provided that this factor is known in advance or can be estimated at testing time. As shown by [Kittler *et al.*, 1998], both rules are approximations to the full posterior probability of the class given all feature streams, obtained by making the assumption that the features are conditionally independent given the class. Kittler notes that while this assumption is likely to be violated in practice, the approximation still gives good results experimentally. [Bilmes and Kirchhoff, 2000] derive the graphical models representing the explicit assumptions made by the product and sum combination rules.

2.6 Connections and Comparisons

Our work lies at the intersection of object recognition, image retrieval, word sense disambiguation and multimodal reference resolution. In image retrieval, an image collection is searched to find images matching a user's query either by text keyword, by low-level image content similarity, or by high-level semantic content. The visual sense model we introduce in Chapter 5 can be considered a high-level CBIR method that takes a text keyword as query and organizes the results by word sense. Although, in this dissertation, we work with image datasets collected by web search, in principle, the method can be applied to any collection of images associated with text, such as news articles, magazines, photo sharing sites. The only modification required would be adding an initial step of keyword-based indexing of the images, similar to Google or Yahoo.

Previous CBIR approaches either focused mostly on simple visual concepts that can be described by uniform colors and textures, or applied supervised learning to a small number of more complex categories [Liu *et al.*, 2007]. On the other hand, our method learns models of arbitrary nouns. The combination of image and text features is used in some web retrieval methods (e.g. [James *et al.*, 2003]), however, our work is focused not on instance-based image retrieval, but on *category-level* modeling.

The Corr-LDA model and others proposed for caption data are relevant to our problem, however, to the best of our knowledge, they have not been extended to deal

with free-text web page contexts. Models of such annotated data often assume that there is a correspondence between each image region and a word in the caption. The focus is on predicting words, which serve as category labels, based on image content. In our case, the goal is to predict a category label based on all of the words in the text context.

Word disambiguation is a classic problem, however, it has only recently been posed for images associated with words. Most previous work is concerned with images paired with human-generated captions (e.g. [Barnard and Johnson, 2005]). In the area of image sense disambiguation applied to keyword-based web image search results, the approach of Loeff et al. [Loeff *et al.*, 2006] attempts to separate the image and text pairs by clustering in both domains, however, since it only performs clustering, it cannot associate a sense label with an image. One of the major contributions of this thesis is the ability to compute the likelihood of a particular word sense for a given web image.

A recent trend in object recognition research is the exploitation of the large number of images available through image search engines to aid in the construction of object category models. The models developed in this work can be used for automated dataset construction for an open vocabulary of objects, with varying degrees of supervision. In Chapter 4, a few examples labeled by the user are required to select the desired object, in Chapter 5 the word sense must be specified, and in Chapter 6, we present a fully automatic method for retrieving only the "physical object" senses of words.

In comparison with existing approaches to web-based dataset construction, our method is the only one that can automatically deal with polysemous words. Reranking with a category-independent text classifier and then with an image classifier trained on the top-ranked results proposed by Schroff et al. [Schroff *et al.*, 2007] fails to deal with polysemous words, since their top-ranked results would likely include several senses. However, theirs is a *category-independent* text classifier and does not learn which words are predictive of a specific category. Another unsupervised approach by Fergus et al. [Fergus *et al.*, 2005] also relies on the top-ranked images

obtained by translating queries to several different languages to select the foreground sense.

Similar to our model, many approaches to dataset construction utilize latent topic models. Berg et. al.[Berg and Forsyth, 2006] discover topics using Latent Dirichlet Allocation (LDA) [Blei *et al.*, 2003] in the text domain. However, their method requires manual intervention by the user to sort the clusters found by LDA into positive and negative for each category. Also, the clusters are learned unimodally, i.e. on words alone, rather than on both words and images.

In multimodal HCI applications, the problem of deixis has inspired research on using images, gesture and speech to resolve references to objects in virtual reality games [Kaiser *et al.*, 2003], to teach a robot about objects in the environment [Haasch *et al.*, 2005], and to learn the meanings of words [Roy *et al.*, 2002]. In Chapter 7, we propose a method that uses both speech and gesture classifiers to resolve object references. In contrast to the virtual reality systems, our method targets the user's real-world environment, where the object identity is unknown. Speech-based object recognition in such environments is explored in the BIRON project [Haasch *et al.*, 2005], however 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 correctly recognized words. Also, 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 focus of Roy and co-authors is on language learning and understanding, whereas our interest is in improving visual recognition. Specifically, we are interested in a realistic object categorization task, and on disambiguating among many arbitrary categories using prior visual models.

3

Data

This chapter describes the datasets used for evaluation in this thesis. In Chapters 4, 5 and 6, we evaluate image retrieval and classification on several image/text datasets collected by querying an internet search engine for a keyword. These are described in Section 3.1 below. In Chapter 7, we evaluate multimodal object classification on images from a standard object recognition dataset (Caltech101 [Fei-Fei *et al.*, 2007]), paired with speech utterances. This dataset is described in Section 3.2. For an overview of existing image-only benchmark data used in the object recognition literature see Chapter 2.1.2.

3.1 Web Image and Text Datasets

The main goals of this work are 1) to create a method that can automatically construct visual models of arbitrary objects using unlabeled results of internet search engines, 2) to build sense-specific models for polysemous words, and 3) to explore the use of image and speech for multimodal object resolution. In order to evaluate the proposed methods on the first two of the above tasks, we need a dataset of images labeled with sense-disambiguated labels. In addition, to evaluate the web image/text retrieval method, we need a sense-labeled dataset of web image search results.

At the time of this writing, several labeled image datasets are available to researchers for this purpose. However, very few of them contain either categories corresponding to polysemous words with sense-disambiguated labels, or image/html data harvested from the web. ImageNet [Deng *et al.*, 2009] (see Chapter 2.1.2) contains sense-disambiguated labels for a large number of concepts, which makes it a good dataset for evaluating our unsupervised method for creating sense-specific classifiers. Unfortunately, ImageNet was released after the experimental portion of this dissertation was completed, therefore testing on that particular dataset remains a direction for future work.

Several of the authors dealing with retrieval of clean images from web search results have collected and annotated datasets of such results and made them available for research. Fergus et al. [Fergus *et al.*, 2005] published a dataset of images downloaded from the web for the keywords 'airplane', 'cars rear', 'face', 'guitar', 'leopard', 'motorbike' and 'wristwatch'. However, this dataset included neither multiple sense labels nor the originating page sources. Similarly, the authors of "Animals on the Web" [Berg and Forsyth, 2006] have made available a dataset for several animal categories that only includes the images and single-sense labels.

On the other hand, the UIUC-ISD dataset released by Loeff et al. [Loeff *et al.*, 2006] contains both the originating page source and the images. It contains polysemous words and human-generated image labels for multiple senses of each word. We use the UIUC-ISD dataset to evaluate sense-specific image retrieval in Chapter 5 and

concrete sense retrieval in Chapter 6. In addition, we collected two datasets that are similar to UIUC-ISD, one containing five polysemous words (MIT-ISD), and one containing ten words describing common office objects (MIT-OFFICE). All three datasets were collected automatically by issuing queries to the Yahoo Image Search engine and downloading the returned images and corresponding HTML web pages. We will now describe each of the datasets in more detail.

3.1.1 UIUC-ISD

The UIUC Image Sense Disambiguation (UIUC-ISD) dataset was collected at UIUC by Loeff et al. [Loeff *et al.*, 2006] for the purpose of evaluating image sense disambiguation on images collected via query search. [1] Three basic query terms were used: BASS, CRANE and SQUASH. The WordNet definitions of these words are shown in Appendix A, Table A.1.

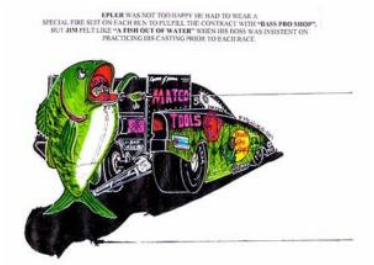
Internet search engines typically limit the total number of returned results per query; in the case of Yahoo, the limit is 1000 images. In our experience, after dead links and corrupted images have been eliminated from the raw results, an average of 700-800 valid image/html pairs can be obtained from one query. To increase corpus size, the authors of UIUC-ISD used supplemental query terms for each word. The search terms selected were those related to each concrete sense of each query (e.g. for CRANE-4, they used "construction cranes", "whooping crane", etc.) Note that these search terms required a human to propose relevant phrases for each word sense. While this would no doubt increase the precision of the corpus, we are interested in methods that do not require any human intervention beyond specifying the word or, at most, the WordNet synset. We therefore exclude the results of manually-generated queries from our experiments.

The images in UIUC-ISD were labeled by human annotators, with several senses labeled for each word. Although the authors made no mention of consulting any dictionary definitions, the annotated senses were common visual meanings of the words

[1] The UIUC-ISD dataset and its complete description can be obtained at http://visionpc.cs.uiuc.edu/isd/index.html

and coincided with the following WordNet synsets: BASS-7 (musical instrument), BASS-8 (fish), CRANE-4 (lift), CRANE-5 (bird), SQUASH-1 (plant), SQUASH-2 (vegetable) and SQUASH-3 (game). The annotators used four different labels for each sense: *core*, *related*, *unrelated* and *people*. Examples of each label are shown in Figure 3-1(a). The *related* label was used when the image seemed related to the core sense of the query but did not depict a core-sense object (e.g. an image of a squash blossom returned for the query "squash vegetable"). The *people* label was used for unrelated images depicting faces or a crowd, which occur frequently due to the people-centric way in which users tend to take pictures. In all of the experiments we merge *unrelated* and *people* labels into a single *unrelated* category.

(a) core



(b) related

(c) unrelated

Figure 3-1: Labeling of BASS-8 (fish) sense in the UIUC-ISD dataset.

3.1.2 MIT-ISD

While very similar to the UIUC-ISD dataset, the MIT Image Sense Disambiguation (MIT-ISD) dataset was collected independently of that effort and differs in several respects. The query terms used were: BASS, FACE, MOUSE, SPEAKER and WATCH. Each of the words has anywhere from 4 to 13 senses in WordNet, shown in Appendix

A, Table A.1. The images were labeled by a human annotator with one or more concrete WordNet synsets for each word. The annotated synsets were: BASS-7 (musical instrument), BASS-8 (fish), FACE-1 (human face), MOUSE-1 (rodent), MOUSE-4 (device), SPEAKER-2 (loudspeaker) and WATCH-1 (timepiece). The annotator was familiar with the dictionary definitions of the senses. Only the full-resolution image and none of the corresponding webpage text was shown to the annotator. The annotation interface made it possible to review and change labels.

For this dataset, each concrete sense was labeled as either *core*, *related*, or *unrelated*. The *core* label was used to indicate that the core sense is present in the image, takes up a significant portion of the center of the image, and is easily recognizable by a human. Images where the core-sense object is relatively small or significantly occluded were labeled as *related*. Examples of each label are shown in Figure 3-1.

3.1.3 MIT-OFFICE

One of the main motivations behind this work is to enable interactive systems to recognize a wide variety of common objects. For example, one application is an assistant robot that can recognize objects typically found in a home or office. To evaluate the ability of our method to construct training datasets for such objects, we collected the MIT-OFFICE dataset. Ten common office object names were used as queries to the search engine: CELLPHONE, FORK, HAMMER, KEYBOARD, MUG, PLIERS, SCISSORS, STAPLER, TELEPHONE, WATCH. In Appendix A, Table A.1, we show the definitions of these words found in WordNet.

The OFFICE dataset was labeled by human annotators as before, however, there was only one core sense labeled per query. For example, for KEYBOARD, only the KEYBOARD-1 sense (device consisting of a set of keys) was labeled as the *core* category, although the KEYBOARD-2 sense (a key holder) can also be considered a concrete visual sense. The senses used in labeling were selected to represent the single most common "office object" meaning of each word. Images where the object was prominent and easily recognizable was labeled *core*. Images where the object was

too small, occluded, or where related items were depicted were labeled *unrelated*.

3.2 Speech and Image Dataset

As mentioned at the start of this chapter, one of the goals of this work is to explore the combination of images and speech for object reference resolution in an interactive system, such as an assistant robot. We are particularly interested in whether or not incorporating image features would benefit speech-based recognition of object names. There are many important questions in designing such a system, including how to extract the part of user's speech containing object references, and how to find image frames that contain the object the user is referring to. However, in these initial explorations, we are mostly interested in the effect of combining the two modalities on classification performance, so we assume that the object name has been extracted from speech and the image frame has been provided.

To evaluate our method, we seek a dataset of images of object paired with spoken utterances describing those objects. In particular, we would like to use realistic images, as well as recordings of real users describing the objects depicted in those images. Since we are not aware of any publicly available databases that contain images paired with spoken descriptions, we collected our own. In particular, we augmented a subset of an existing 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 fusion method in Chapter 7.

3.2.1 Image Dataset

As mentioned in Chapter 2.1.2, most publicly available image databases suitable for category-level recognition contain very few object categories. We chose to use the *Caltech101* database, because it contains a relatively large variety of categories, and because it is a standard benchmark in the object recognition field on which several methods have demonstrated a high level of performance. The latter consideration is an important one: while objects such as those in the MIT-OFFICE dataset are

preferable, current object recognition methods do not necessarily work well on those categories, as they have been optimized to perform well on standard benchmarks such as cars, faces and Catech101.

As implied by the name, the database has a total of 101 categories, with an average of 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 in a stereotypical pose. Sample images from each of the 101 categories are shown in Figure 3-2.

3.2.2 Speech Collection

We augmented a subset of the images in Caltech101 with spoken utterances recorded in our lab, producing a set of image-utterance pairs. To limit the vocabulary to 101 names, users were prompted with the exact name of each object. We chose the set of names 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 is shown as captions in Figure 3-2.

A total of six subjects participated in the data collection, four male and two female, all native speakers of American English. Each subject was presented with two images from each category in the image test set, and asked to say the exact object name for each image. Thus, across all six speakers, there were 12 image-utterance pairs for each category, for a total of 1212 image-utterance pairs in the dataset. The reader might question the necessity of showing the image to the subjects, in addition to prompting them with the object's name. One reason for this is that some names are homonyms (e.g., here "bass" refers to the fish, not the musical instrument). Another reason is to make the experience resemble the real scenario of teaching a robot about objects.

The recording process took place in a quiet office, on a laptop computer, using its built-in microphone. The resulting audio was very clean, with a high signal-to-noise ratio. To simulate more realistic home environments, we added "cocktail party"

noise to the original waveforms, using increasingly lower signal-to-noise ratios (SNRs): 10db, 4db, 0db, and -4db.

Figure 3-2: Sample images from the *Caltech-101* database. The category name used in our experiments is shown at the top of each image.

4

Semi-Supervised Image Sense Disambiguation

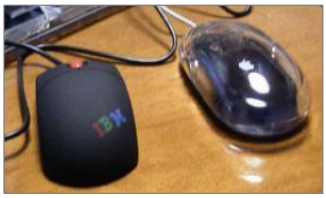
In this chapter, we develop a semi-supervised method for disambiguating the sense of an image returned by web keyword search. In contrast to previous work, our method learns sense-specific topics in both the image features and the context information contained in the words surrounding the image link. Learning is done largely on unlabeled data, with a few labeled examples provided by the user. We propose and evaluate two strategies for combining the text and image features. The evaluation is focused specifically on polysemous nouns, i.e. nouns with multiple dictionary meanings. We compare our multimodal text- and image-based topic learning approach to unimodal baselines. We also explore the effects of learning a distribution over the hidden topics rather than choosing a single best-fitting topic.

4.1 Introduction

The problem of *image sense disambiguation (ISD)* is the problem of categorizing an image indexed by a word, e.g. "mouse", to reflect the specific sense of that word depicted in the image, e.g. "computer mouse". Although the indexing is often assumed to be performed by an internet search engine, the problem applies more generally to any corpus of text illustrated by images. However, the unstructured nature of web pages makes ISD more difficult. There are several applications of ISD. One is in re-organizing the results of text-query image search by sense for better display to the user. Another application is in the exploitation of the large number of images available through image search engines to aid in the construction of object category models.

The ISD problem occurs because the precision of the images returned from web text-query search is often poor [Collins *et al.*, 2008]. This is not surprising, given that web search engines rely mostly on simple text cues, such as the presence of the query word in the filename of the image, and not on image content [Cai *et al.*, 2004]. The query word is not always a reliable cue, since words can have variable meaning depending on the context. The ambiguity is even higher for polysemous words, i.e. words with multiple dictionary meanings. For example, the first page of results returned by an image search engine for the query "mouse" might contain multiple senses of the word, such as: "computer mouse", "four-legged mouse", "Mickey Mouse".

Existing solutions to the ISD problem include iterative co-training [Feng *et al.*, 2004], image-only bootstrapping from labeled data [Li *et al.*, 2007] , clustering the unlabeled images into coherent components [Loeff *et al.*, 2006, Fergus *et al.*, 2005], and iterative re-ranking [Schroff *et al.*, 2007]. These approaches vary in the amount and quality of supervision required from the user. Iterative co-training and bootstrapping methods require the presence of a seed labeled dataset in order to train the initial classifiers. In the case when such a dataset is not available, several methods use the top-ranked results returned by the text-query search as positive examples ([Fergus *et al.*, 2005, Schroff *et al.*, 2007]). However, this approach fails in the case



…mac pc wheel parallels virtualpc mac pc wheel wheel mouse mac wheel mouse wheel wheel mouse mac mouse control mac control control alt control macbook powerbook parallels parallels plan macbook parallels parallels preference keyboard mouse right click alt click plan shift parallels plan alt shift ok alt shift virtualpc powerbook virtualpc plan virtualpc control click shift click control click shift contrl click plan…

…light compatible windows xp nt standard ps mouse port works surface glass home alien dvds alien pictures alien videos ufo tv schedule links beta tester new product alienvideo net like invite beta tester ufo mouse beta tester try new mouse order today free shipping days return product completely satisfied performance ufo mouse ufo left beta test make order today miss beta tester send product test required beta testers encouraged send feedback hand comfort…

Figure 4-1: Two images and corresponding text contexts returned for the query MOUSE.

of polysemous words, for which the top-ranked results are likely to include several senses.

The semi-supervised paradigm of extracting coherent clusters from the noisy search results ([Fergus *et al.*, 2005, Berg and Forsyth, 2006]) has the desirable property that it takes advantage of the large number of unlabeled data and thus requires fewer labeled examples. A small amount of validation data is used to select the cluster that aligns the best with the positive class. For example, Berg et al. [Berg and Forsyth, 2006] discover topics using LDA [Blei *et al.*, 2003] in the *text* domain, and then use them to cluster the images. However, their method requires manual intervention by the user to sort the clusters into positive and negative for each category. Also, the clusters are learned unimodally, on words alone, rather than on both words and images.

Fergus et al. [Fergus *et al.*, 2005] also perform topic clustering, but they do it based on just the image features, ignoring the text contained in the corresponding web pages. The limitation of learning distances in a single space is that often the representation in that space is insufficient to distinguish semantically different examples. For example, an image of a white bird in a blue sky cannot be distinguished from a similarly-colored image of an airplane using color histogram features. Furthermore, because the models learn only clusters similar to the labeled data, generalization to different types of objects in the same class is restricted. For example, if the labeled

images do not include computer mice shaped like a UFO (see Figure 4-1), the model may reject them as being too dissimilar from the training set.

Both of these problems can be mitigated by incorporating multiple complementary modalities into the learning process. In this chapter, we extend the semi-supervised learning approach of Fergus et al. [Fergus *et al.*, 2005] to use both text and image features to learn same-sense clusters from unlabeled data. The advantages of our multimodal scheme over text-only and image-only clustering are 1) better robustness and 2) better generalization. Better robustness comes from the fact that, when the two types of features are conditionally independent, the classifiers' errors tend to be uncorrelated, and they are able to correct each other's mistakes. Better generalization comes from not restricting the classifier to instances that are only visually similar to the labeled set. In the example above, our method can take advantage of the text context of a UFO mouse being similar to text in the labeled data, even if the image is not.

This chapter is organized as follows: Section 4.2 presents an image sense disambiguation method that is based on learning hidden topics in a corpus of images and associated text. Implementation details of image and text processing are given in Section 4.3. Experiments on a dataset collected using polysemous nouns as web queries are presented in Section 4.4. Section 4.4.1 gives a qualitative evaluation of the learned topics, and Section 4.4.2 shows that combining text and image information can result in improved performance over using either modality alone. We conclude the chapter with a discussion of the results in Section 4.5.

4.2 Approach

The textual context information typically available in web search scenarios consists of HTML source code with the image link embedded in it. We use the term *text context* to refer to the 100 or so words surrounding the image link in the code, as well as text in the relevant tags, such as the ALT tag. The text context potentially contains useful information about the particular meaning of the word: For example, a

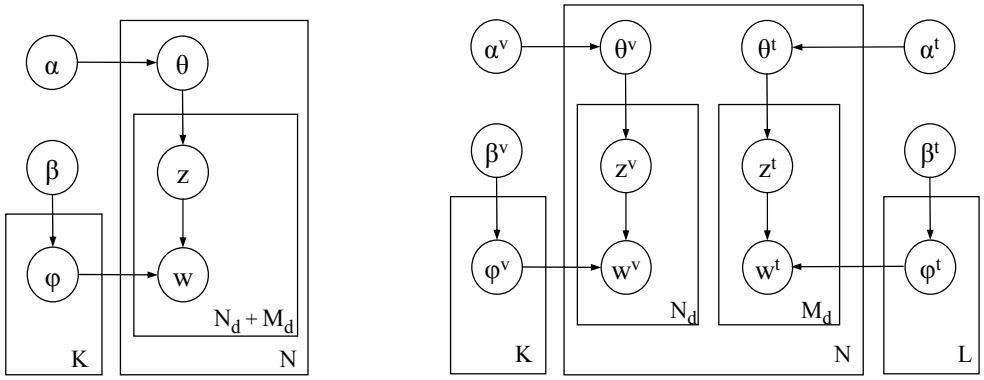
web page selling watches might surround watch images with text describing the dial, strap, brand, and so forth, whereas another web page posting an image with "tornado watch" in the filename might contain weather-related words. Such text, however, can be highly unconstrained and noisy (see example in Figure 4-1). One of the goals of this chapter is to evaluate how useful such noisy text is for classifying the *visual* sense depicted in the image.

We follow the approach of treating both images and text contexts as unordered discrete observations, or "bags of words". In the case of images, the "words" are local patch descriptors quantized into visual words using a codebook. From now on, unless specified otherwise, we will use the terms "document" and "words" to refer interchangeably to text contexts and words, and to images and visual words.

The main idea behind our approach is to extract cohesive components from the unlabeled image/text data that correspond to different meanings, or senses, of the query word. This approach was proposed by Fergus et al. [Fergus *et al.*, 2005] for learning visual components. Here we also use the text context to guide the formation of components, and explore different ways of combining modalities. There are two classic approaches to modality fusion: late fusion (i.e. combination at the classifier level) and early fusion (i.e. concatenation of features). We explore both approaches within the framework of a latent Dirichlet allocation (LDA) model (see Chapter 2.2.2). First, we describe the early-fusion model.

4.2.1 Early-Fusion Model

We construct an early-fusion multimodal LDA model in a straightforward manner: we treat visual and text words identically and allow a topic to represent distributions over both types of words. The intuition is that, similarly to how unimodal LDA discovers clusters of words that frequently co-occur in documents, multimodal LDA may discover that certain visual words tend to co-occur with certain text words. The graphical model is shown in Figure 4-2(a). It is identical to unimodal LDA, except that, if $|V|$ is the size of the visual vocabulary and $|T|$ of the text vocabulary, the word variable w takes values $w \in \{1, ..., |V| + |T|\}$. Also, the total number of word



(a) Early (b) Late

Figure 4-2: Graphical models of (a) early-fusion LDA and (b) late-fusion LDA. The superscripts in (b) refer to the modality: (v)isual, (t)ext.

observations in document d is $M_d + N_d$, where M_d is the number of visual words in the image, and N_d is the number of words in the text context. Once the features have been combined in this way, inference proceeds as in regular LDA.

The advantage of this model is that it is straightforward to implement and can potentially capture correlations between visual and textual manifestations of senses. The disadvantage is that, as with most early-fusion approaches, the increase in the dimensionality of the data means that more data may be required to properly train the model. Also, as it is described above, the model does not account for the possibility of mismatched amounts of visual and text words. In fact, as we will show in Section 4.4.2, when words in one modality greatly outnumber words in the other modality, the latter do not have as much influence on the formation of topics.

4.2.2 Late-Fusion Model

Our late-fusion LDA method works by fitting a separate LDA model in each modality, independently of each other. Figure 4-2(b) shows the corresponding graphical model. The idea is to delay the interaction of the visual and text topics until the time of classification, when the category label is inferred based on the multimodal image-text pair. The label is inferred using late fusion of two classifers: the classifier based on text topics and the classifier based on image topics. This amounts to assuming that

the distributions of image and text words are conditionally independent of each other, given the category label, which is a reasonable assumption in our case. There are two advantages to this model: 1) it does not suffer from the unbalanced word numbers in the two views like the early-fusion model does, and 2) it allows the modalities to be weighted differently during classification.

4.2.3 Classification Algorithm

The input to the algorithm consists of image and web page pairs retrieved by text-query image search. The data is split into a large number of unlabeled pairs and a small validation set. The user is asked to label each image in the validation set as either positive (depicting the correct sense) or negative. The output is a categorical label $y \in \{-1, +1\}$ for each unlabeled pair.

First, a hidden topic model – either early-fusion LDA, or late-fusion LDA – is learned on the unlabeled data. Then, following the approach of [Fergus *et al.*, 2005], we classify the validation set with each of the resulting K topics. For a given topic j, a document d is classified as positive if the probability of the topic, $P(z = j|d) = \theta_d^j$ (estimated using Eq. (2.4)), exceeds a certain threshold. Varying this threshold allows one to compute a recall-precision curve over the validation dataset. Finally, a single topic is chosen that has the best performance (in terms of the area under the recall-precision curve). The other topics are assumed to represent the negative class (background).

In the case of the text-only, image-only, and early-fusion models, a single best topic j is chosen to represent the positive sense. In the case of the late-fusion model, an image topic j and a text topic k are chosen independently. The probabilities of the two topics are combined as follows:

$$P(y = +1|d, j, k) \equiv \lambda\, P(z^v = j|d) + (1 - \lambda)\, P(z^t = k|d) \tag{4.1}$$

where $y \in \{-1, +1\}$ is the category label, d is the document consisting of the text/image pair, and $\lambda \in [0, 1]$ is a weight that controls how much each modality

contributes to the final decision. The weight can be set by cross-validation or fixed.

While LDA has been shown to discover coherent topics in text documents (e.g. politics, science, etc.), the assumption that a single visual-word topic will capture the positive class is rarely justified in practice, except for very simple cases [Larlus and Jurie, 2009]. This is because, unlike words in a language, which represent whole concepts, visual words correspond to small texture patches and are not nearly as descriptive. Therefore, our use of a single topic to embody the inlier sense may be too restrictive. To explore and alternative, we also propose a multi-topic classifier. The hidden topics are discovered as before, but, instead of committing to one topic, we learn a Dirichlet distribution over the topic proportions θ given the positive class:

$$P(\theta|y=+1) = Dir(\theta|\alpha, y=+1), \tag{4.2}$$

The probability of the inlier sense given a document is then

$$P(y=+1|d) \equiv P(y=+1|\theta_d) \propto P(\theta_d|y=+1)P(y=+1). \tag{4.3}$$

A similar distribution is learned for the negative class. A sample with topic distribution θ_d is labeled as positive if the following decision rule holds true

$$log(\frac{P(\theta_d|y=+1)}{P(\theta_d|y=-1}))>0. \tag{4.4}$$

and negative otherwise. For the late-fusion multi-topic classifier, the decision values of the text and image models are combined in a similar manner to the single-topic case.

4.3 Features

This section describes the processing of the dataset required to extract image and text words. The following processing steps were applied to the raw dataset: Image/HTML pairs that contained unreachable URLs and/or corrupted images were removed from

the dataset. Furthermore, pairs for which the algorithm failed to extract a text context were removed. This mostly happened when the link to the image could not be located in the HTML document, such as when the original webpage was changed or removed.

Text Bag-of-Words. To extract text context words for each image, the image link was located automatically in the corresponding HTML page. All HTML tags were removed, and the remaining text was tokenized. A standard stop-word list of common English words was applied (adding a few domain-specific words like "jpg"), followed by a Porter stemmer [Porter, 1988] to extract word stems (e.g. "us" from "use"). Word stems that appeared only once and the actual query word stem were pruned. Finally, all word tokens in a 100-token window surrounding the location of the image link were extracted. The resulting vocabulary size (per query word) ranged between 3500 and 4500 words. Each text context was represented as a histogram of counts for each word in the vocabulary.

Image Descriptors. All images were resized to 300 pixels in width and converted to grayscale. Two types of local feature points were detected in the image: edge features [Fergus *et al.*, 2005] and scale-invariant salient points. In our experiments, we found that using both types of feature points boosted classification performance relative to using just one type. To detect edge features, we first performed Canny edge detection, and then sampled a fixed number of points along the edges from a distribution proportional to edge strength. The scales of the local regions around points were sampled uniformly from the range of 10-50 pixels. To detect scale-invariant salient points, we used the Harris-Laplace [Mikolajczyk and Schmid, 2004] detector with the lowest strength threshold set to 10. Altogether, 400 edge points and approximately the same number of Harris-Laplace points were detected per image. A 128-dimensional SIFT descriptor was used to describe the patch surrounding each interest point.

Image Bag-of-Words. After extracting a set of interest point descriptors for each image, vector quantization into visual words was performed. A codebook of size 800 was constructed by k-means clustering on a randomly chosen subset of the

Table 4.1: The 10 most-likely words in LDA topics learned for MOUSE.

1	2	3	4	5	6	7	8
watch	page	pad	mice	home	mouse	gif	optical
mickey	custom	mouse	like	cat	frame	index	usb
click	products	item	new	pictures	thumbnail	size	wireless
disney	animals	product	posted	site	pet	gallery	use
contact	order	great	web	new	user	description	logitech
band	home	quality	news	make	add	directory	keyboard
price	tcp	resolution	carlpc	picture	img	modified	design
photos	world	gift	store	copyright	advertise	la	laser
information	accessories	note	posts	search	jmk	di	creative
samuel	october	personalized	pm	photo	simian	photo	hand

database (300 images per query). All images were converted to histograms over the resulting visual words. To be precise, the "visual words" correspond to the cluster centers (codewords) of the codebook. Note that no spatial or color information was included in the image representation in these experiments.

4.4 Experiments

In this section, we outline the experimental setup and results of image sense disambiguation. The dataset used for evaluation in this chapter is the MIT-ISD dataset described in Chapter 3.1.2. In all of the following experiments, Gibbs sampling was carried out using the Matlab Topic Modeling Toolbox [Steyvers and Griffiths,].

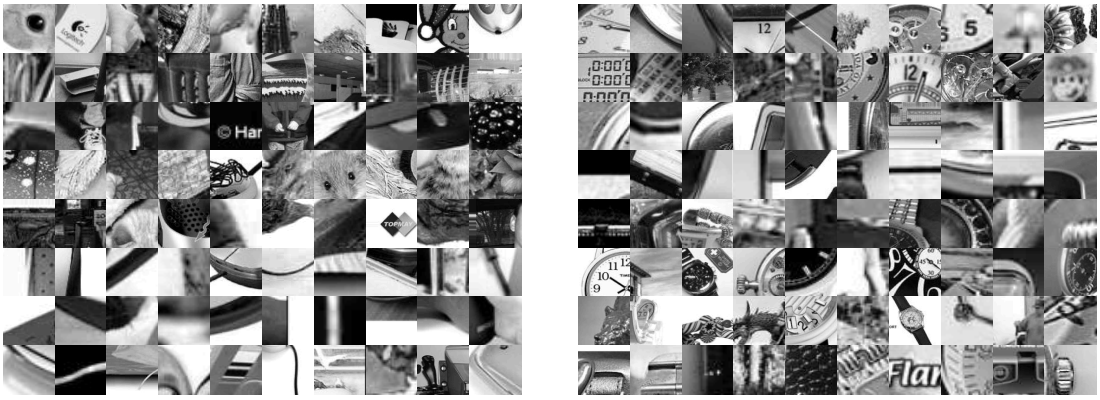
4.4.1 Qualitative Analysis of Learned Topics

Having learned the hidden topics in an unsupervised fashion, one might ask: how meaningful are they? Do the text words align with our intuitive understanding of different meanings of each query word? An example of the text topics learned using late-fusion LDA with $K = 8$ for the query MOUSE is shown in Table 4.1. Each column shows the top ten most likely words in the distribution for one topic. Upon

analysis of the topics, they do seem to correspond to the common meanings of the query words that can be used to describe images. For MOUSE, topic 8 has words like "optical", "usb", etc., suggestive of computer devices. Other learned topics seem to do with Mickey Mouse watches, cat and mouse, and some more general background topics.

For the other query words in the MIT-ISD dataset, the topics (not shown here) also tend to align with different senses of the word. For BASS, the topics seem to have to do with either fishing or guitars. For FACE, the different meanings are not as clear, but some words in topics are suggestive of possible meanings: "rock face", "funny face", "smiley face". Some of the topics align not with word meanings, but rather background topics that have to do with types of web pages on which one might find images. For example, there seem to be blog topics (with words "blog", "comment", "post"), e-commerce topics ("price", "usd"), and image gallery topics ("home", "gallery").

We also examine the learned visual topics. Figure 4-3 visualizes the topics learned on the image data of the MOUSE and WATCH query words. The other queries' results are similar. Each row shows the 10 most likely visual word examples for one topic, in order of decreasing likelihood from left to right. The visual word examples were picked at random from the dataset: For each visual word (codeword), a random image was chosen that contains a patch assigned to that visual word, and this patch was shown in the figure. One must keep in mind that different-looking patches can be assigned to the same word. In general, it is more difficult to analyze the visual topics, as the patches do not have an easily deduced meaning. However, one observation we can make is that some of the mouse topics prefer simple edges, suggesting computer mice on white backgrounds, while others prefer more natural textures, suggesting animals in outdoor scenes, fur, etc. For WATCH, the third topic from the top seems to be picking out rounded parts of the watch face.



(a) MOUSE (b) WATCH

Figure 4-3: Each row shows the ten most likely visual words for one topic.

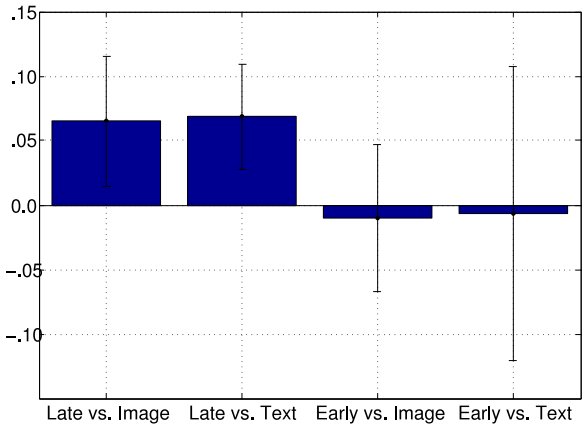


Figure 4-4: **Early vs. Late fusion:** Average difference in area under the RPC between each fused model and text- and image-only models is shown.

4.4.2 Image Sense Disambiguation

In this section, we evaluate the proposed multimodal sense classifiers and compare them to baselines that use text or image features alone. The following issues are investigated: 1) whether combining modalities using either early or late fusion LDA benefits classification, 2) whether multi-topic classification is better than single-topic classification, and 3) the effect of fixed model parameters λ and K on performance.

Evaluation metric. The evaluation task is to classify the unlabeled image/text pairs as either depicting the core sense or not. Classification of a single core sense was evaluated for each keyword: BASS-8 (fish), FACE-13 (human face), MOUSE-4 (rodent), SPEAKER-2 (loudspeaker), WATCH-1 (timepiece). As mentioned in

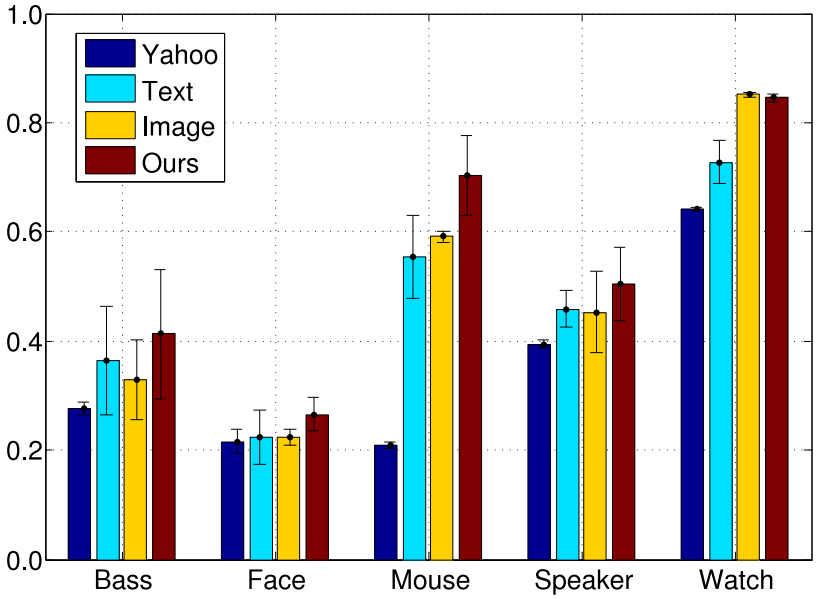
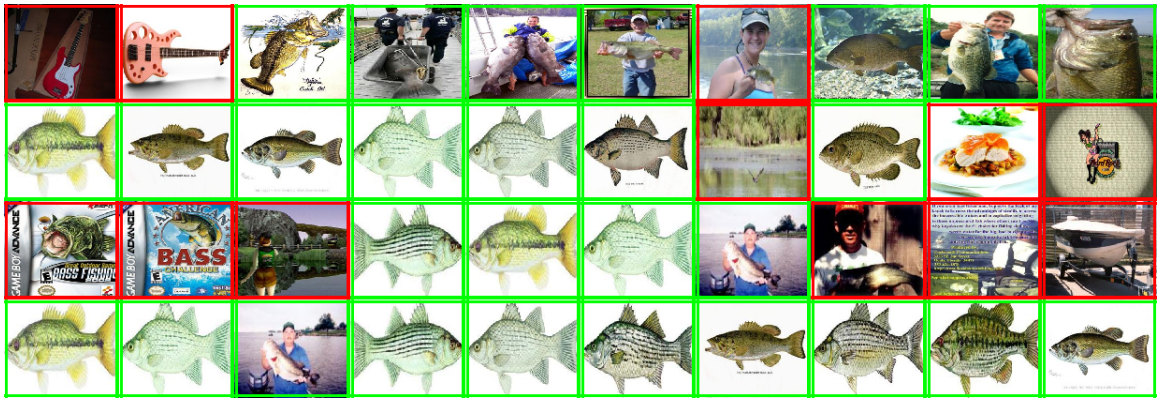


Figure 4-5: **Results for each keyword:** Area under the RPC is shown for several methods ("ours" means the late-fusion model).

Chapter 3.1.2, each sense in MIT-ISD was labeled as *core*, *related*, or *unrelated*. In the following experiments, only *core* labels are mapped to positive labels, and *related* and *unrelated* are grouped into the negative class. To quantitatively compare models, we conduct a ten-fold classification experiment by randomly splitting the data into a 20-pair validation set and an unlabeled test set. The labels are held out when training the LDA models on the unlabeled data, and only used to compute recall and precision. The unlabeled data are assigned labels by thresholding $P(y = +1|d)$. Precision (the number of true positives divided by the total number of samples labeled as positive) and recall (the number of true positives divided by the total number of positive samples) are computed at each threshold. The area under the resulting recall-precision graph, which corresponds to average precision (the higher, the better), is used as the evaluation metric.

Early vs. Late LDA. Figure 4-4 shows the relative improvement obtained by the early- and late-fusion classifiers over the baseline text-only and image-only classifiers. The plot shows the absolute difference in area under the recall-precision curve (RPC), averaged over trials and over the five words in MIT-ISD. While the late-fusion method obtains a significant improvement over both baselines, the early-fusion method does



(a) BASS-8

(b) MOUSE-4

Figure 4-6: Images top-ranked by Yahoo (first row), the image-only method (second row), the text-only method (third row) and the late-fusion method (fourth row).

not. The failure of the early-fusion model may be caused by the fact that it does not properly normalize for the disparity in the number of words across modalities (there are more than eight times the number of visual words than there are text words in each pair). Since visual words have more influence in the model, its performance is essentially limited by the image-only baseline. In the rest of the experiments, we only consider the late-fusion model.

Multimodal vs. Unimodal. Figure 4-5 shows the actual areas under the RPC for our method (late-fusion classifier) and several baselines. For each keyword, the average area is reported with the error bars showing standard deviation. Our model consistently improves upon the original Yahoo recall-precision curve, which means it is able to achieve a higher precision of the true sense, based on the input features. The text-only and image-only models generally improve on the original Yahoo precision,

except for the case of the FACE query, where the target sense cannot be distinguished based on either the image or the text context. Our multimodal approach tends to achieve either the best of the text-only or image-only performances, or improve on both. Overall, it is better than using either modality alone.

Re-ranking Example. Figure 4-6 illustrates the benefit we get from text and image data fusion. Each row corresponds to a particular method's re-ranking of unlabeled images, for queries BASS and MOUSE. The positive categories are core senses BASS-8 (fish) and the MOUSE-4 (computer device). Images whose ground truth labels are negative are outlined in red. For each query, the top row shows the original top ten Yahoo images. The next three rows show the ten most likely images for the positive class: The second row shows images for the best image-only topic, the third row for the best text-only topic, and the fourth row for the combination of the best text and image topics using late fusion. The original Yahoo results contain images of mixed meanings of each word (music and fishing, device and animal). The image topic tends to cluster images with similar features together, but makes mistakes (e.g. cooked fish). The text-only classifier does well at selecting a single meaning of the word, but the images are not always representative of the visual object (e.g. people fishing, boats). On the other hand, when combined in the late-fusion model, the topics tend to correct each others mistakes.

Single- vs. multi-topic model. We also compare a classifier that picks a single best topic to one that learns a distribution over topic proportions to represent the positive class. Figure 4-7 shows the area under RPC for both types of classifier, using the late-fusion model. For three out of five queries, the multi-topic model outperforms the single-topic approach; for the other two queries it performs comparably.

Parameter Selection. In the above experiments, the number of topics was set to $K = 8$, and the text model weight to $\lambda = 0.5$. Picking the K that gives the best results on the small validation set is not a good idea, since this tends to favor large K and over fit. Fergus, et. al. [Fergus *et al.*, 2005] found that $K = 8$ worked well, and that for K greater than about 10 the validation set was less able to predict a good topic. Figure 4-8 shows a similar finding for our late-fusion model. The dashed

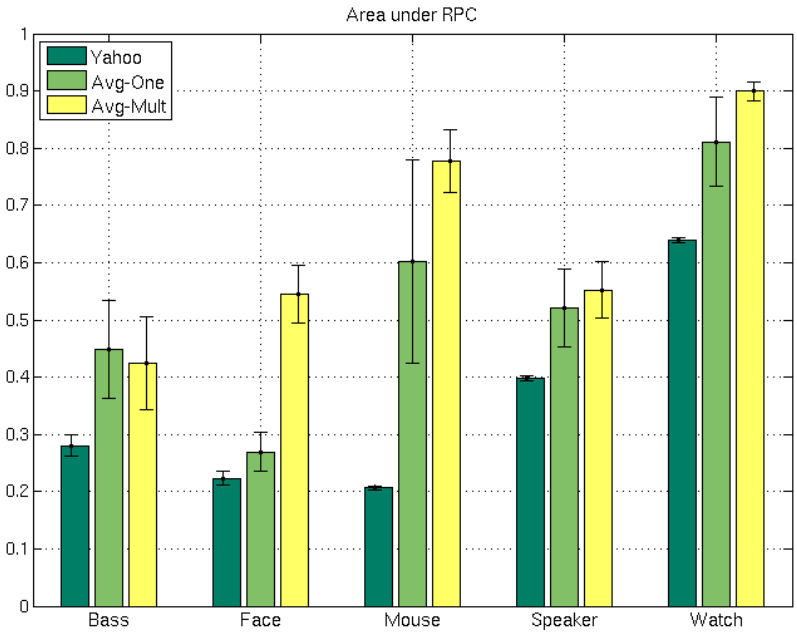


Figure 4-7: **Single- vs. multi-topic model:** Average area under RPC is shown.

curve shows the performance of the classifier using the best topics chosen on the test data, averaged over categories. The solid line shows the best validation topics' performance, which peaks around $K = 8$ and starts to diverge from the that of test set-picked topic for $K > 8$. Figure 4-9 shows the effect of varying λ, evaluated on the test set. We see that, on average across categories, the fused classifier improves upon unimodal classifiers ($\lambda = 0$ and $\lambda = 1$) in the range between 0.2 and 0.9; the value of 0.5, which assumes that text and image features are equally important, is in that range.

4.5 Discussion

In this chapter, we have argued that a multimodal approach to category learning from web image search engines is advantageous because it leads to increased robustness and generalization. We proposed two LDA-based models of text context and image data, one based on concatenation of features, and another based on combination of classifier decisions. Both models learn a hidden topic space on the large available

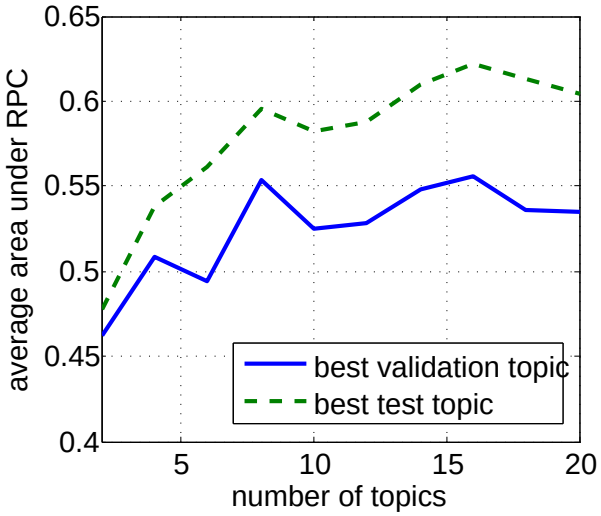


Figure 4-8: **Varying K:** As the number of topics K increases, performance (on test data) of the best topic chosen on the validation set diverges from that of the best topic chosen on the test data.

unlabeled dataset obtained by image search, and select the best topic or combination of topics based on performance on a small number of hand-labeled examples. We have compared the proposed multimodal methods to the original search engine retrieval and to the unimodal (text- and image-only) baselines.

The evaluation has shown that classifiers based on text alone sometimes outperform image-based classifiers, however, neither is a clear winner across all categories in our dataset. However, the late-fusion approach benefited from the redundancy of the text and image features, allowing the unimodal clusters to correct each other's mistakes and outperforming all baselines on average across categories. We also found that our early-fusion LDA approach suffers from an imbalance in the number of text and visual words, the latter outnumbering the former by a factor of eight or greater.

Our semi-supervised classifier is based on LDA, which has been shown to be effective at learning coherent and meaningful topics in both text and image domains. However, in principle, other clustering methods could be used in our general framework, as long as they are able to provide well-defined posterior probabilities of the cluster given the data. A direction for future work is to attempt to address the aforementioned normalization problem in the early-fusion model, and to explore other

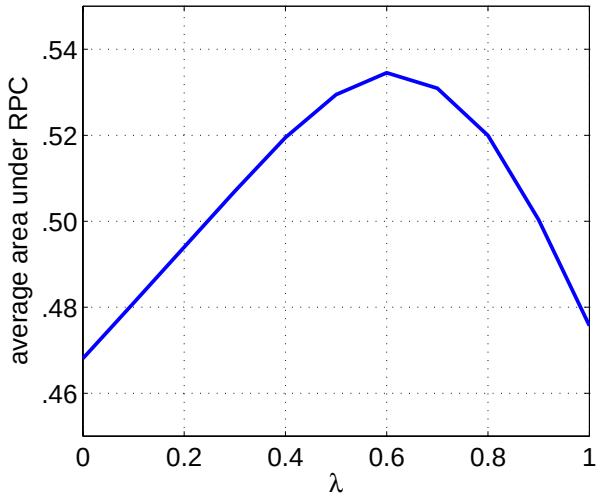


Figure 4-9: **Varying lambda:** effect of the text model weight on test set performance.

LDA-based models of text and image data for this problem, such as a modified version of Corr-LDA [Blei and Jordan, 2003] that does not assume a correspondence between each visual region and a text word. Another avenue for future work is a more principled approach to selecting the number of clusters, perhaps by using infinite mixture models, such as the HDP [Teh *et al.*, 2003].

In this chapter, we learn hidden topics using LDA directly on the words surrounding the images. However, while the resulting topics were often aligned along sense boundaries, the approach suffers from over-fitting, due to the irregular quality and low quantity of the data. Often, the text context is just a short text fragment, such as "fishing with friends" for an image returned for the query "bass". The topics learned on this data tend to over-fit, learning unusual words that are specific to websites/webpages over-represented in the search results. In the next chapter, we propose an alternative that alleviates this problem.

One major drawback to the methods outlined in this chapter is the requirement of labeled examples to learn the inlier topics. This most likely means asking the user to manually several label examples of the object he/she is looking for. In Chapter 7, we propose a robust way of using speech to label objects in the interactive scenario. However, many applications would benefit from a method that can learn visual models in a way that does not require any supervision from the user. One of the benefits of

incorporating text-based models of image sense into our method is the possibility of using text-based ontologies to reduce the amount of required supervision. In the next chapter, we introduce a method that does just that, and apply it to the problems of image sense disambiguation and object recognition.

5

A Dictionary Model of Image Sense

In this chapter, we continue to address the problems of automatic image sense disambiguation and automatic dataset construction. In contrast to the previous chapter, we no longer require that the user manually label examples of the desired sense. Instead, we learn image sense models in an unsupervised way, using existing text-based knowledge repositories of word senses to guide the learning process. [1]

5.1 Introduction

In the previous chapter, we introduced a method that can be used to retrieve images of a word's visual sense, if both the word and a few labeled examples of the desired visual sense are available. However, manually labeled images are costly to obtain.

[1] Portions of this chapter were published in [Saenko and Darrell, 2008]

One example where image labels are difficult to obtain is the "home tour" scenario (see Chapter 2). The robotic assistant may have access to a user's *spoken description* of an object, but not necessarily to image examples to ground the word sense. In this case, word sense disambiguation may still be possible based on the results of speech recognition. For example, given the utterance "I am going to read a book, bring me my glasses," the system may infer that "glasses" refers to "spectacles", and not "drinking glasses", as indicated by the collocations "read" and "book".

The goal of this chapter is to develop an *unsupervised* approach to ISD, where the only information required besides the word is the word sense. In addition to the interactive scenario described above, where the word sense may be inferred from language, such an approach can be applied more generally in any scenario where a list of word senses is known. For example, one might cluster image search results by dictionary sense, or build sense-specific visual models. Although word senses do not always co-incide with physical objects, for now, we will assume that the desired sense is indeed a visual one. In the next chapter, we will address the problem of identifying non-visual senses automatically.

Existing unsupervised approaches to automatic dataset construction attempt to filter out images unrelated to the desired object, but do not directly address polysemy. Often the search query is tailored by the researcher in an effort to narrow down the category, e.g. "computer mouse", and polysemous words are generally avoided. Most existing approaches rely on a labeled seed set of inlier-sense images to initialize bootstrapping or to select the right cluster [Li *et al.*, 2007, Fergus *et al.*, 2005, Berg and Forsyth, 2006]. The unsupervised approach of Schroff et al. [Schroff *et al.*, 2007] bootstraps a classifier from the top-ranked images returned by a search engine, with the assumption that they have higher precision for the desired object. However, for polysemous words, the top-ranked results are likely to include several senses.

As shown in the previous chapter, the words surrounding web images indexed by a polysemous keyword can be a rich source of information about the sense of that word. But how can we learn a model of sense without any labeled images? One idea is to use repositories of word sense knowledge, such as online dictionaries and ontologies,

•S: (n) **mouse** (any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)
•*direct hyponym* / *full hyponym*
  •S: (n) house mouse, Mus musculus (brownish-grey Old World mouse now a common household pest worldwide)
  •S: (n) harvest mouse, Micromyx minutus (small reddish-brown Eurasian mouse inhabiting e.g. cornfields)
  •S: (n) field mouse, fieldmouse (any nocturnal Old World mouse of the genus Apodemus inhabiting woods and fields and gardens)
  •S: (n) nude mouse (a mouse with a genetic defect that prevents them from growing hair and also prevents them from immunologically rejecting human cells and tissues; widely used in preclinical trials)
  •S: (n) wood mouse (any of various New World woodland mice)
•***direct hypernym*** / *inherited hypernym* / *sister term*
  •S: (n) rodent, gnawer (relatively small placental mammals having a single pair of constantly growing incisor teeth specialized for gnawing)

Figure 5-1: WordNet entry for one sense of the word "mouse", including its hyponyms and hypernyms.

to ground visual senses. In its most general form, a dictionary is a list of entries that define the senses of each word in a language. An example entry for one of the senses of the word "mouse" contained in the WordNet dictionary[2] is shown in Table 5-1. The WordNet ontology also contains lexical information such as example sentences, *synonyms*, *hyponyms* and *hypernyms*. A potential method of learning sense models in an unsupervised way would be to use this information as a word observations to seed a probabilistic model of each sense, as it is defined by entries in the dictionary. This is similar to the unsupervised WSD algorithm proposed by Yarowsky [Yarowsky, 1995].

In this chapter, we introduce such an unsupervised dictionary-based ISD method, one that specifically takes word sense into account. The only input to the algorithm is a list of words (such as all English nouns, for example) and their dictionary entries. The proposed method is multimodal, in that it uses both web search images and the text surrounding them in the document in which they are embedded. It learns a model of the word sense using an electronic dictionary together with a large amount of unlabeled text. The use of a dictionary is key because it frees us from needing a labeled set of images to ground the sense. The model can retrieve images of a specific sense from the mixed-sense image collection, and the re-ranked images can be used as training data for an sense-specific object classifier. The resulting classifier can predict not just the word that best describes a novel image, but also the correct *meaning* of

[2]See `http://wordnetweb.princeton.edu/perl/webwn?s=mouse`

that word.

The rest of the chapter is structured as follows: Section 5.2 describes the method, Section 5.4 the features, and Section 5.5 experimental evaluation, which includes both sense retrieval from web search results and classification of unseen images. Section 5.6 concludes the chapter.

5.2 Approach

Since this work is concerned with objects rather than actions, we restrict ourselves to the noun senses of words. As in standard word sense disambiguation (WSD) methods, we make a one-sense-per-document assumption [Yarowsky, 1995], and rely on words co-occurring with the image in the HTML document to disambiguate that sense. However, image links are not guaranteed to be surrounded by grammatical sentences, which makes it difficult to extract structured features such as part-of-speech tags and apply traditional WSD methods. We therefore once again take a bag-of-words approach, using all available words near the image link to evaluate the probability of the sense. This is accomplished by a latent topic model that predicts which words are likely for the sense. The proposed method consists of three main steps:

1. discovering latent topics associated with a word,

2. learning a topic-based probabilistic model of dictionary senses, and

3. using the model to construct sense-specific image classifiers.

We will now describe each step of our method, which we refer to as *W*eb *I*mage *S*ense *D*icti*O*nary *M*odel, abbreviated *WISDOM*.

5.2.1 Latent Text Space

As mentioned above, our goal is to learn a probabilistic model of words that are likely for a particular word sense. While dictionary entries contain examples of such words, they are usually limited in size and can only provide coverage of a very small

portion of the input word space. A possible supervised approach is to learn on a sense-disambiguated corpus, one that is used in traditional WSD, and apply the learned model to web data. However, based on observation, image text contexts are sufficiently different from such corpora, and a better approach might be to learn on data obtained from the web. We can extend the limited coverage of dictionary entries by leveraging the fact that, while sense-disambiguated examples of web text are rare, unlabeled web text related to the word in question is abundant. Such text can be obtained, for example, by extracting the word context of each occurrence of the keyword in web pages returned by a search engine.

The first step of the *WISDOM* algorithm is thus to use a large collection of text related to the word to learn coherent dimensions. The hope is that these dimensions will fall along different senses or uses of the word. Several existing techniques could be used to discover latent dimensions in documents consisting of bags-of-words. Here, as in the previous chapter, we use latent Dirichlet allocation. In Chapter 2.2.2 we gave a review of LDA; we now briefly review the notation for the convenience of the reader.

Each document is modeled as a mixture of topics $z \in \{1, ..., K\}$. A given collection of M documents, each containing a bag of N_d words, is assumed to be generated by the following process: First, we sample the parameters ϕ^j of a multinomial distribution over words from a Dirichlet prior with parameter β for each topic $j = 1, ..., K$. Then, for each document d, we sample the parameters θ_d of a multinomial distribution over topics from a Dirichlet prior with parameter α. Finally, for each word token i, we choose a topic z_i from the multinomial θ_d, and then pick a word w_i from the multinomial ϕ^{z_i}.

In Chapter 4, we learned text topics on a corpus consisting of the words surrounding the images. Such text contexts are often short, sometimes consisting only of a text fragment. The irregularity of text contexts is compounded by the ad-hoc structure of web pages, with unrelated text often appearing close to the image file. Furthermore, search engines limit the number of images they return for a query, typically to 1000 results, which further limits the amount of available data. While the resulting topics

were often aligned with senses, the approach suffered from over-fitting, due to the irregular quality and low quantity of the data (see Chapter 4.5). As an example, we refer back to Table 4.1, which showed sample topics learned from image contexts returned for the query MOUSE. While some of the topics contain likely words that are indicative of the "rodent" sense (e.g. "animal", "pet"), it is difficult to assign a single topic that is clearly aligned with that sense. Compare this to topic 8, which is is strongly suggestive of the "computer device" sense of MOUSE.

To allieviate the aforementioned overfitting problem, we create an additional unlabeled dataset of text-only web pages. This is done by sending the basic keyword as a query to a *web* search engine, such as Google or Yahoo. We then fit an LDA model to the obtained dataset and use the resulting topic distributions to constructing a model of the dictionary senses, as described in the next section.

5.2.2 A Text Model Based on WordNet

Table 5.1: WordNet semantic relations included in *WISDOM*.

Relation	Definition	Example	Included?
synonym	Y is a synonym of X if they have very similar meanings	bug is a synonym of germ	✓
hypernym	Y is a hypernym of X if every X is a (kind of) Y	canine is a hypernym of dog	1st-level
hyponym	Y is a hyponym of X if every Y is a (kind of) X	dog is a hyponym of canine	✓
coordinate term	Y is a coordinate term of X if X and Y share a hypernym	wolf is a coordinate term of dog	
holonym	Y is a holonym of X if X is a part of Y	building is a holonym of window	✓
meronym	Y is a meronym of X if Y is a part of X	window is a meronym of building	✓

WISDOM uses the limited text available in dictionary entries to relate each sense to latent topics formed as described above. Here, we will present a version that uses the WordNet lexical database, although a different dictionary, thesaurus or ontology

can be used in its place. The advantage of WordNet is that it provides *semantic relations* between words. Word senses are grouped into *synsets*, or sets of synonyms, each of which represents a single concept. Examples of synsets are given in Appendix A, Table A.1. Various relations link the concepts represented by synsets. In the case of nouns, these are "part-whole" relationships and "is-a" relationships. Table 5.1 shows the noun relations that are used in *WISDOM* to access additional content for a sense entry. We exclude hypernyms higher than the first level because they are very general concepts. Coordinate terms are excluded because they contain entire classes of concepts, such as all canines.

Given a query word with WordNet senses $s \in \{1, 2, ...S\}$, the definition and semantically related items are concatenated together to produce the sense entry. For instance, for sense $s = 1$ of PLIERS, this entry consists of the synonyms "pair of pliers, plyers", the definition "a gripping hand tool with two hinged arms and (usually) serrated jaws", the first-level hypernym "hand tool", the hyponyms "locking pliers, needlenose pliers, pump-type pliers, rib joint pliers, slip-joint pliers", and the meronym "jaw". We denote the bag-of-words extracted from an entry with the variable $e_s = \{w_1, w_2, ..., w_{E_s}\}$, where E_s is the total number of words in the entry. Next, we outline two alternative formulations of generative models of image sense based on such entries.

Mixture-of-Multinomials Model. The first model we propose is a generative model of image contexts based on the mixture-of-multinomials model. Each text context belongs to a single sense, which generates a topic, which in turn generates the words. The assumption here is that the observed words are independent of the sense given the underlying topic. A text context document d consisting of words $\{w_1, w_2, ..., w_{N_d}\}$ is generated as follows:

1. pick a sense $s_d \in 1, ..., S$ from a prior distribution $P(s)$,
2. select a topic $z_d \in 1, ..., K$ from $P(z|s_d)$, a multinomial distribution with parameter η_{s_d},
3. for each word token i, choose a word w_i from $P(w|z_d)$, a multinomial with

parameter ϕ^{z_d}.

The probability of a document is

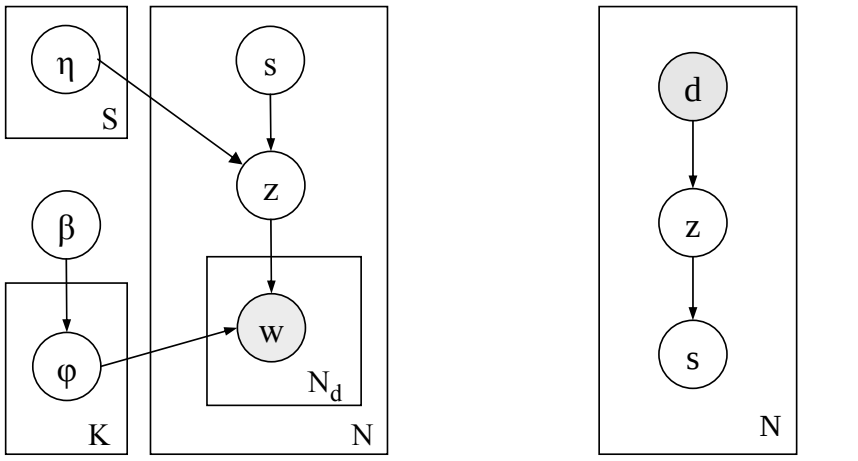
$$P(w_1, ..., w_{N_d}) = \sum_{h=1}^{S} \sum_{j=1}^{K} \prod_{i=1}^{N_D} P(w_i|z = j)P(z = j|s = h)P(s = h) \tag{5.1}$$

The graphical model is shown in Figure 5-2(a). Note that, in this paradigm, there is a single topic per document. This is a limiting assumption for complex documents such as entire webpages, but for the relatively short (100 or so words) text contexts, it is not unrealistic. The parameter η of the distribution of topics for each sense is fixed for the entire corpus.

LDA-Factor Model. The mixture-of-multinomials sense model is intuitive, but suffers from a major drawback. It treats the topic proportions inside a document as a fixed parameter and not as a random variable, as it is done in LDA. Without the smoothing provided by the prior, overfitting can occur, especially since many text contexts contain very few words. In an alternative approach, we compute the topic proportions in the text context using LDA, with a $Beta(\alpha)$ prior on the multinomial θ parameter. This model does not explicitly generate words, but rather treats documents d and their topic proportions θ_d as observed variables. Because it states that the sense is independent of the observations conditioned on the latent topic, or factor, we call this model the LDA-Factor model. For a web image with an associated text document $\{w_1, w_2, ..., w_{N_D}\}$, the probability of sense conditioned on that document is given by

$$P(s|d) = \sum_{j=1}^{K} P(s|z = j)P(z = j|d). \tag{5.2}$$

The above requires the distribution of latent topics in the text context, $P(z|d)$, and the probability of the sense given the latent topic, $P(s|z)$. The former is given by the θ_d variable computed by generalizing the LDA model trained on the text-only data to the (unseen) text contexts. Note that, while the LDA model allows multiple topics to be associated with one document, for the purposes of the sense model, a single topic variable is associated with one document, as in the mixture-of-multinomials model.



(a) Mixture-of-Multinomials Sense Model

(b) LDA-Factor Sense Model

Figure 5-2: Graphical representations of the sense models.

Figure 5-2(b) displays the corresponding graphical model.

To obtain $P(s|z)$, we first compute the likelihood of s given latent topic $z = j$ as the average likelihood of words in the entry e_s, or

$$P(z|s) \propto \frac{1}{E_s} \sum_{i=1}^{E_s} P(w_i|z), \tag{5.3}$$

normalized so that it constitues a probability distribution over z. The average word likelihood was found to be a good indicator of how relevant a topic is to a sense. The total word likelihood could be used, but would mean that senses with longer entries dominate. Using Bayes's rule, we obtain

$$P(s|z) = \frac{P(z|s)P(s)}{P(z)} \tag{5.4}$$

The text-only model we have outlined defines the probability of a particular dictionary sense given an image/text pair to be equal to $P(s|d)$. Thus, the model is able to assign sense probabilities to images returned from the search engine, which in turn allows it to group the images according to sense.

5.2.3 Incorporating Image Features

The text-only LDA-Factor model computes $P(s|d)$, where d is the text context. Thus, it does not take into account the image part of the image/text pair. Here, we extend this model to include an image term, which, as we showed in Chapter 4, can potentially provide complementary information. From this point on, we will refer to the text observation as d^t and the image observation as d^i. We call the joint image and text model *WISDOM-2*.

First, we estimate $P(s|d^i)$, or the probability of a sense given an image d^i. Similar to the text-only case, we learn an LDA model consisting of latent topics $v \in \{1, ..., L\}$, using the visual bag-of-words extracted from the unlabeled images in the dataset. The estimated θ variables give $P(v|d^i)$. To compute the conditional probability of a sense given a visual topic, we marginalize the joint $P(s, v)$ across all document pairs $\{d^i, d^t\}$ in the collection

$$P(s|v) = \frac{\sum_{k=1}^{M} P(s|d_k^t)P(v|d_k^i)}{P(v)}. \tag{5.5}$$

Note that the above assumes conditional independence of the sense and the visual topic given the observations. Intuitively, this provides us with an estimate of the collocation of senses with visual topic.

We can now compute the probability of dictionary sense for a novel image d_*^i as:

$$P(s|d_*^i) = \sum_{j=1}^{L} P(s|v=j)P(v=j|d_*^i) \tag{5.6}$$

Finally, the joint text and image model is defined as the combination of the text-space and image-space models via the sum rule,

$$P(s|d^i, d^t) = \lambda P(s|d^i) + (1-\lambda)P(s|d^t) \tag{5.7}$$

Our assumption in using the sum rule is that the combination can be modelled as a mixture of experts, where the features of one modality are independent of sense given the other modality [Bilmes and Kirchhoff, 2000].

5.2.4 Classification of Novel Images

The last step in *WISDOM* uses the sense model learned in the first two steps to generate training data for an image-based classifier. The choice of classifier is not an essential part of the algorithm. We choose to use a discriminative classifier, in particular, a support vector machine (SVM), because of its ability to generalize well in high-dimentional spaces without requiring a lot of training data.

For each particular sense s, the model re-ranks the images according to the probability of that sense, and selects the N highest-ranked examples as positive training data for the SVM. The negative training data is drawn from a "background" class, which in our case is the union of all other objects that we are asked to classify. We represent images as histograms of visual words, which are obtained by detecting local interest points and vector-quantizing their descriptors using a fixed visual vocabulary.

We compare the *WISDOM* classifier with a simple baseline method that attempts to refine the search by automatically generating search terms from the dictionary entry, described in the next section.

5.3 Baseline

A human operator is often able to refine the search by using sense-specific queries, for example, searching for "computer mouse" instead of "mouse". We explore a simple method that does this automatically by generating sense-specific search terms from entries in WordNet. Experimentally, we found that queries consisting of more than two or three terms returned very few image results. Consequently, the terms are generated by appending the polysemous word to its synonyms and first-level hypernyms. Multiple word occurrences are removed. For example, the sense MOUSE-4 has a synonym "computer mouse" and a hypernym "electronic device", which produces the query terms "mouse computer" and "mouse electronic device". An SVM classifier is then trained on the returned images.

Because the terms method must rely on one- to three-word combinations, it can be brittle. Many of the generated search terms are too unnatural and bookish to

retrieve any results (e.g. "percoid bass"). Some retrieve too many unrelated images, such as the term "ticker" used as an alternative to "watch". *WISDOM* overcomes these issues by learning a model of each sense from a large amount of text. Web text is more natural, and is a closer match to the type of text surronding web images than dictionary words are. This makes *WISDOM* more robust, as will be shown in the experimental section.

5.4 Features

When extracting words from web pages, all HTML tags are removed, and the remaining text is tokenized. A standard stop-word list of common English words, plus a few domain-specific words like "jpg", is applied, followed by a Porter stemmer [Porter, 1988]. Words that appear only once and the actual word used as the query are pruned. To extract text context words for an image, the image link is located automatically in the corresponding HTML page. All word tokens in a 100-token window surrounding the location of the image link are extracted. The text vocabulary size used for the sense model ranges between 12K-20K words for different keywords.

To extract image features, all images are resized to 300 pixels in width and converted to grayscale. Two types of local feature points are detected in the image: edge features [Fergus *et al.*, 2005] and scale-invariant salient points. In our experiments, we found that using both types of points boosts classficiation performance relative to using just one type. To detect edge points, we first perform Canny edge detection, and then sample a fixed number of points along the edges from a distribution proportional to edge strength. The scales of the local regions around points are sampled uniformly from the range of 10-50 pixels. To detect scale-invariant salient points, we use the Harris-Laplace [Mikolajczyk and Schmid, 2004] detector with the lowest strength threshold set to 10. Altogether, 400 edge points and approximately the same number of Harris-Laplace points are detected per image. A 128-dimensional SIFT descriptor is used to describe the patch surrounding each interest point. After extracting a bag of interest point descriptors for each image, vector quantization is

performed. A codebook of size 800 is constructed by k-means clustering a randomly chosen subset of the database (300 images per keyword), and all images are converted to histograms over the resulting visual words. To be precise, the "visual words" are the cluster centers (codewords) of the codebook. No spatial information is included in the image representation, but rather it is treated as a bag-of-words.

5.5 Experiments

In this section, we evaluate *WISDOM* on the tasks of ISD and classification of novel images, and compare it to the baseline terms method. The datasets used for evaluation in this chapter are the MIT-ISD and UIUC-ISD datasets described in Chapter 3.1.2. In all of the following experiments, Gibbs sampling was carried out using the Matlab Topic Modeling Toolbox [Steyvers and Griffiths,], and the SVM was implemented using the LIBSVM toolbox [Chang and Lin, 2001].

For the following experiments, we collected two additional sets of unlabeled training data. The first set of data is the images collected using the generated sense-specific search terms to augment the MIT-ISD dataset (see Section 5.3.) This data was used to train the baseline classifier. The second set of data was collected via regular web search, using the original keywords, for both the MIT-ISD and the UIUC-ISD datasets. Bag-of-words data were extracted from the web pages and used to train the text component of *WISDOM*. Table 5.2 shows the size of the additional datasets for MIT-ISD and the distribution of labels.

5.5.1 Qualitative analysis of text topics

First, we examine the learned text topics to gauge the benefit of using a separate corpus of webpages vs. the text contexts of images. Table 5.3 shows the web topics learned for the MOUSE query. When we compare them to the topics in Table 4.1 learned from text contexts, several differences emerge. The main difference is that the web topics are more general than image context topics. Another observation is that web topics seem to consitute better models of different senses of the word. For

Table 5.2: **MIT-ISD additional data:** sizes of the text-only, sense-term, and keyword datasets, and distribution of ground truth sense labels in the keyword dataset.

category	size of datasets			distribution of labels in the keyword dataset	
	text-only	sense-term	keyword	positive (core)	negative (related, unrelated)
Bass	984	357	678	146	532
Face	961	798	756	130	626
Mouse	987	726	768	198	570
Speaker	984	2270	660	235	425
Watch	936	2373	777	512	265

example, topics 1 and 2 are clearly "computer device" topics, and topic 6 is likely a "rodent" topic.

In addition, several of the web topics have to do with scientific research involving mice, judging from the words "gene", "research", "protein", etc. While these topics can help disambiguate the "rodent" sense of the word mouse, they are specialized to a particular area and may not be represented in image search results. In either case, our algorithm is flexible in that it is able to "ignore" irrelevant topics by assigning them a low likelihood.

5.5.2 ISD Using Text Features

The goal of these experiments is to evaluate how well *WISDOM* can distinguish between images depicting the correct visual sense and all the other senses, based only on their text contexts.

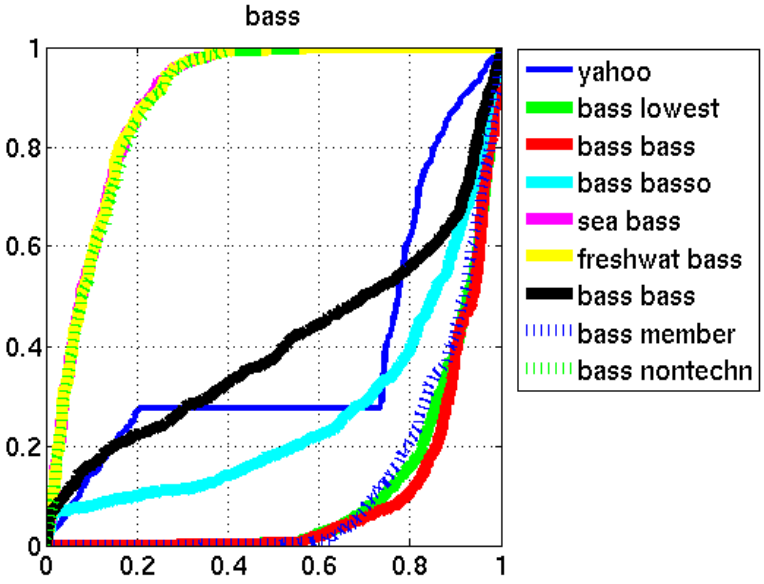
Evaluation metric. The evaluation task is to classify the unlabeled image/text pairs as either depicting the core sense or not. As described in Chapter 3.1.2, senses in evaluation data are labeled as either *core*, *related*, or *unrelated*. In the following experiments, only *core* labels are mapped to positive labels, while *related* and *unrelated* are grouped into the negative class. The labels are held out when training the LDA models on the unlabeled data, and only used in evaluation. The unlabeled

Table 5.3: 20 word stems from 8 LDA topics learned for MOUSE, sorted by decreasing likelihood (top to bottom).

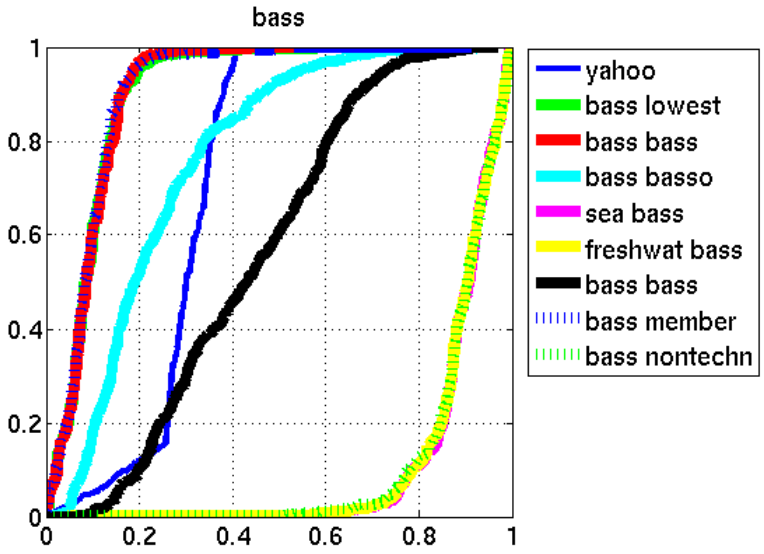
1	2	3	4	5	6	7	8
optic	button	comment	anti	new	mice	cell	gene
pad	us	video	fc	research	post	express	sequenc
product	click	mickei	ciali	mice	trap	abstract	genom
review	window	new	antibodi	anim	just	al	human
usb	user	add	kit	human	like	et	data
wireless	right	blog	purifi	univers	hous	human	us
design	download	site	mg	develop	us	text	map
price	cursor	disnei	rat	speci	make	articl	chromosom
keyboard	left	view	bui	scienc	littl	mice	transcript
laser	devic	music	pe	genet	pm	protein	articl
pc	set	librari	gener	copyright	rat	activ	al
custom	control	email	view	center	place	result	clone
game	appl	http	alpha	servic	sai	us	genet
logitech	softwar	nih	goat	home	rodent	fig	et
microsoft	keyboard	organ	receptor	inform	small	embryo	analysi
technolog	wheel	game	igg	year	cat	promot	region
home	movement	repli	effect	work	look	stem	databas
item	screen	pictur	clone	us	don	site	dna
card	driver	track	affin	write	live	free	primer
us	support	brain	il	includ	thing	cultur	differ

data are assigned labels by thresholding $P(s|d)$. In addition to precision and recall, we also compute the receiver operating characteristic (ROC). The ROC plots the fraction of true positives against the fraction of false positives at each threshold.

Experimental settings. We train a separate text LDA model for each keyword on the web text dataset, setting the number of topics K to 8 in each case. Although this number is roughly equal to the average number of senses for the given keywords, we do not expect nor require each topic to align with one particular sense. In fact, multiple topics can represent the same sense. Rather, we treat K as the dimensionality of the latent space that the model uses to represent senses. While our intuition is that it should be on the order of the number of senses, it can also be set automatically by cross-validation. In our initial experiments, different values of K did not significantly alter the results. We used symmetric Dirichlet priors with scalar hyperparameters $\alpha = 50/K$ and $\beta = 0.01$, which have the effect of smoothing the empirical topic distribution, and 1000 iterations of Gibbs sampling.

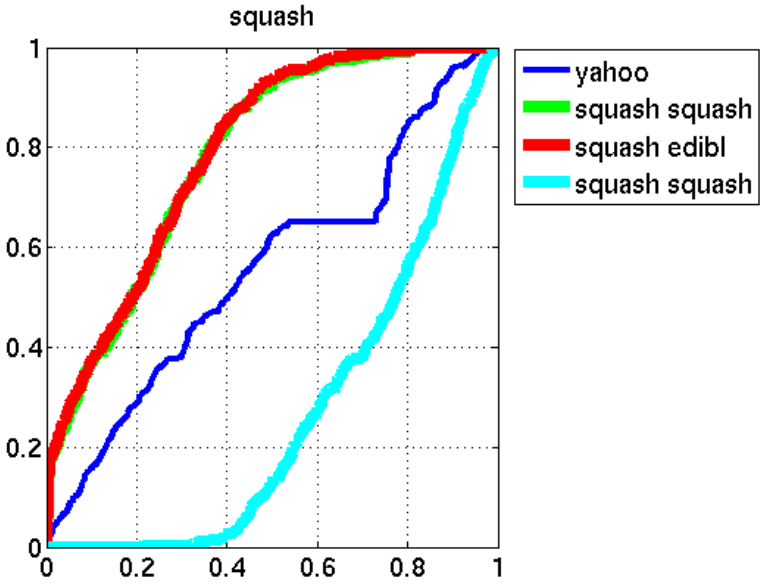


(a) BASS-8 (FISH)

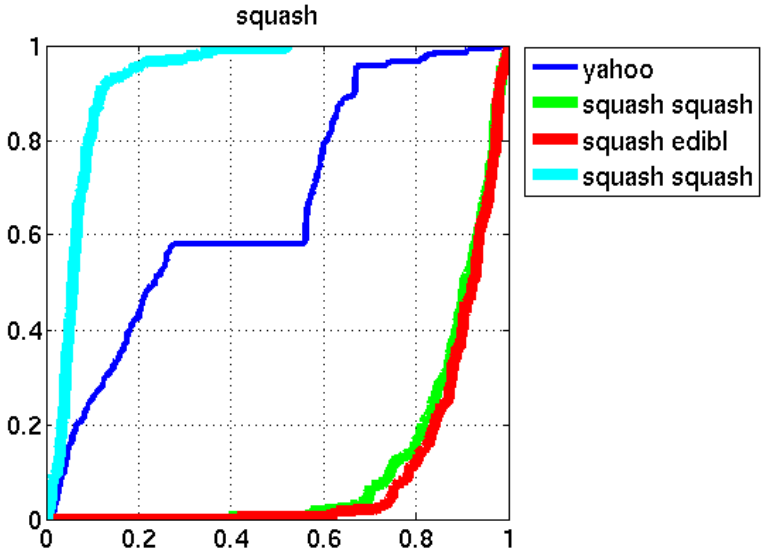


(b) BASS-7 (MUSIC. INSTR.)

Figure 5-3: **Retrieval** of BASS senses in UIUC-ISD. ROCs are shown for the original Yahoo search ranks (blue) and *WISDOM* model of all possible WordNet senses.



(a) SQUASH-2 (VEGETABLE)



(b) SQUASH-3 (GAME)

Figure 5-4: **Retrieval** of SQUASH senses in UIUC-ISD. ROCs are shown for the original Yahoo search ranks (blue) and *WISDOM* model of all possible WordNet senses.

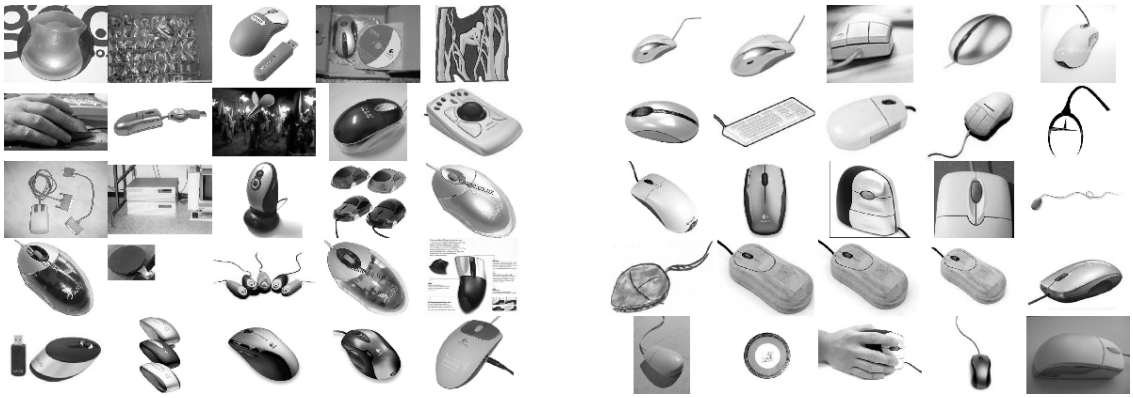
Results. We evaluate the retrieval performance of ground truth image senses using *WISDOM* models of different dictionary senses. We only show results for the LDA-Factor model, as it tends to slightly outperform the Mixture-of-Multinomials model. Figure 5-3 shows the resulting ROCs for BASS in UIUC-ISD, computed by thresholding $P(s|d)$ and scoring it against labels of BASS-8 (fish) and BASS-7 (musical instrument) senses. The eight models corresponding to each of the senses of the keyword "bass" are shown using different colors and line types. The solid blue curve is the ROC obtained using the original Yahoo retrieval order. Figure 5-4 shows the same for SQUASH.

These results demonstrate that the WordNet-based model retrieves far more *core*-label images than the original search engine order. This is important for improving the precision of training data used in the classification step. Of course, not all sense generate good ROCs, as is expected. For BASS and SQUASH, as for several other keywords, there are multiple dictionary definitions that the model is too coarse to distinguish. For example, all three senses "sea bass", "freshwater bass" and "non-technical bass" are about the same atidentifying the fish sense of bass. In the rest of the evaluation, we do not make such fine-grained distinctions, but simply choose the WordNet sense that applies most generally.

As a side note, in interactive applications, the human user can specify the intended sense of the word by providing an extra keyword, such as by saying or typing "bass fish". The correct dictionary sense can then be selected by evaluating the probability of the extra keyword under each sense model, and choosing the highest-scoring one.

5.5.3 ISD Using Text and Image Features

Next, we evaluate the full text and image model on retrieval of all image senses in the two datasets. First, we train a separate text LDA model and a separate image LDA model for each word in the dataset, setting $K = 8$ each. This was done for the text model so that the number of latent text topics would roughly equal to the number of senses. In the image domain, it is less clear what the number of topics should be. Ideally, each topic would coincide with a visually coherent class of images



(a) Text Model (b) Image Model

Figure 5-5: The top 25 images returned by the text and the image models for MOUSE-4 (device).

all belonging to the same sense. In practice, because images of an object class on the web are extremely varied, multiple visual clusters are needed to encompass a single visual category. Our experiments have shown that the model is relatively insensitive to values of this parameter in the range of $K = 8, ..., 32$. As before, we use symmetric Dirichlet priors with scalar hyperparameters $\alpha = 50/K$ and $\beta = 0.01$ and 1000 iterations of Gibbs sampling.

Figure 5-5 shows the images that were assigned the highest probability for MOUSE-4 (computer device) by the text-only model $P(s|d^t)$ (Figure 5-5(a)), and by the image-only model $P(s|d^i)$ (Figure 5-5(a)). We observe that both models return high-precision results, but somewhat different and complementary types of images. The image model's results are more visually coherent, while the text model's results are more visually varied, which is what we would expect to happen.

The recall-precision curves (RPCs) of isolated senses are shown in Figure 5-6 for *WISDOM-2* (green curves) and the Yahoo rank order (blue curves). RPCs are computed for each labeled sense in the MIT and UIUC-ISD datasets by thresholding $P(s|d^i, d^t)$ (Eq. 5.7). For example, the top leftmost plot shows retrieval of BASS-7 (musical instrument). These results demonstrate that we are able to greatly improve the retrieval of each concrete sense compared to the search engine, without any supervision. That said, this comparison is somewhat unfair to the search engine as

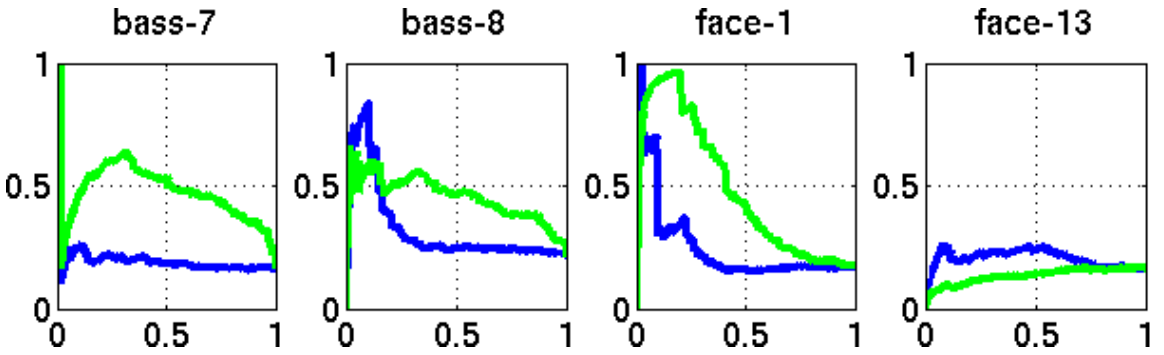
our method has one piece of knowledge the engine does not – the dictionary sense number.

Our model does fail to retrieve one sense, FACE-13. There happened to be quite a few mountain cliff images in the Yahoo results, prompting the labeler to mark them as a separate sense of FACE. However, in WordNet FACE-13 is defined only as "a vertical surface", a very vague and terse definition. This is a highly visually ambiguous sense, one that could potentially include a very diverse class of images, and not just the cliff faces that were labeled in MIT-ISD. In addition, none of the LDA topics for FACE seem to align with this meaning of the word, which likely caused the poor performance. Had there been a strong "mountain cliff" topic, then the model might have overcome the terseness of the definition.

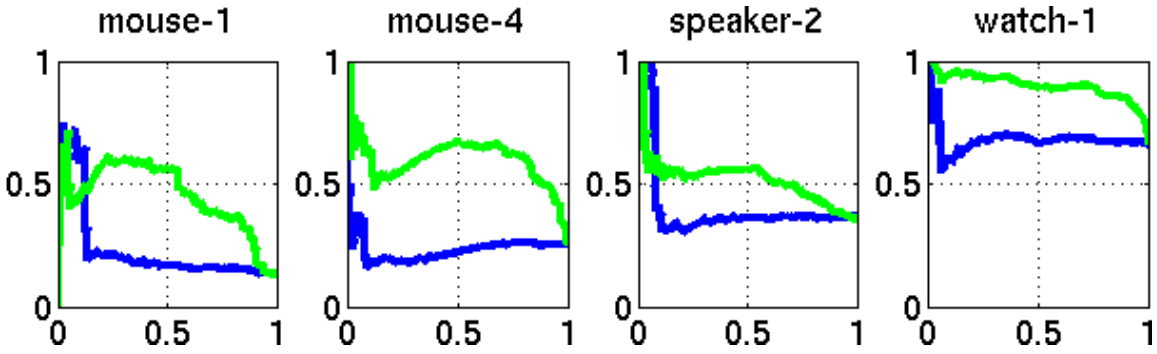
5.5.4 Classifying Unseen Images

The goal of these experiments is to evaluate *WISDOM* on an object classification task where only a novel image is provided as input. The evaluation is carried out on the MIT-ISD dataset.

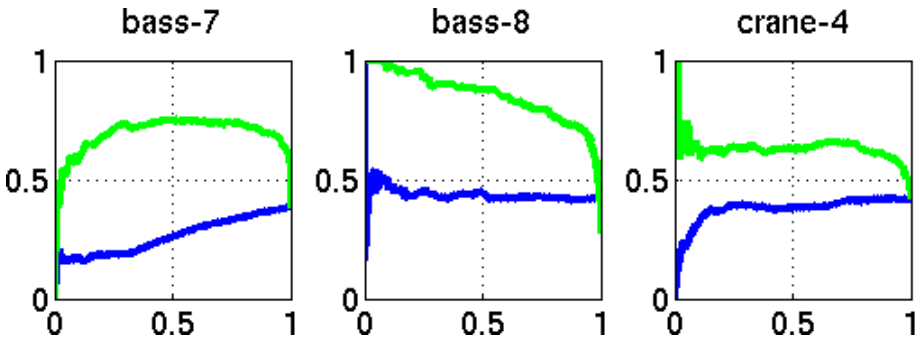
Experimental Settings. We train classifiers for five objects corresponding to the following image senses: BASS-8 (fish), FACE-1 (human face), MOUSE-4 (device), SPEAKER-2 (loudspeaker) and WATCH-1 (timepiece). The classifiers are binary, assigning a positive label to the correct sense and a negative label to incorrect senses and all other objects. The top N unlabeled images ranked by the sense model are selected as positive training images. The unlabeled image pool consists of both the keyword and the sense-term datasets. N negative images are chosen at random from positive data for all other keywords. A binary SVM with an RBF kernel is trained on the image features, with the C and γ parameters chosen by four-fold cross-validation. The baseline algorithm is trained on a random sample of N images retrieved using the automatically generated sense-specific query terms. Recall that the terms were generated from word combinations extracted from the target sense definition (see Section 5.3. Training on the first N images returned by Yahoo did not qualitatively change the results.



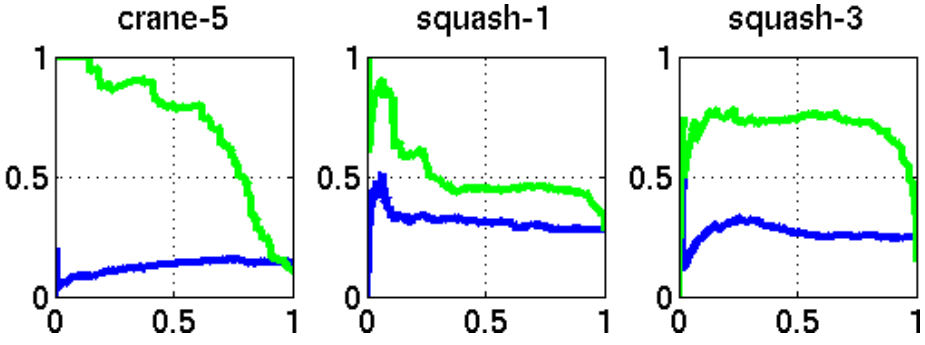
(a) MIT data



(b) MIT data



(c) UIUC data



(d) UIUC data

Figure 5-6: Retrieval of isolated senses (core labels) using *WISDOM-2* on two datasets.

We evaluate the method on two test cases. In the first case, the negative class is composed only of images from the other words. We refer to this as the 1-SENSE test set. In the second case, the negative class also includes other senses of the target word. For example, if we are testing classification of MOUSE-4 (device), the negative class includes images of any sense of BASS, FACE, SPEAKER and WATCH, as well as "animal mouse", "Mickey Mouse" and other non-MOUSE-4 images in the MOUSE dataset. We refer to this as the MIX-SENSE test set.

Results. Figures 5-7,5-8 and 5-9 compare the classification accuracy of *WISDOM* to the baseline auto-terms classifier. Average accuracy across ten trials with different random splits into train and test sets is shown for each object. Figure 5-7 shows results on 1-SENSE and 5-8 on MIX-SENSE, with N equal to 250. Figure 5-9 shows 1-SENSE results averaged over all five categories, at different numbers of training images N. In both test cases, our dictionary method significantly improves on the baseline algorithm. As the per-object results show, we do much better for three of the five objects, and comparably for the other two. One explanation why we do not see a large improvement in the latter cases is that the automatically generated sense-specific search terms happened to return relatively high-precision images. However, in the other three cases, the term generation fails while our model is still able to capture the dictionary sense.

5.6 Discussion

While labeled examples of image senses are rare, an abundance of human knowledge about word senses exists in the form of electronic dictionaries, encyclopedias, and semantic databases. In this chapter, we introduced a way to harness that knowledge in creating unsupervised models of image sense. To the best of our knowledge, this is the first use of a dictionary in either web-based image retrieval or in object recognition. Another key feature of the algorithm is the use of a large amount of unlabeled text available through keyword search on the web in to learn a generative model of sense. The approach is unsupervised, requiring no labeled images of the desired object, and

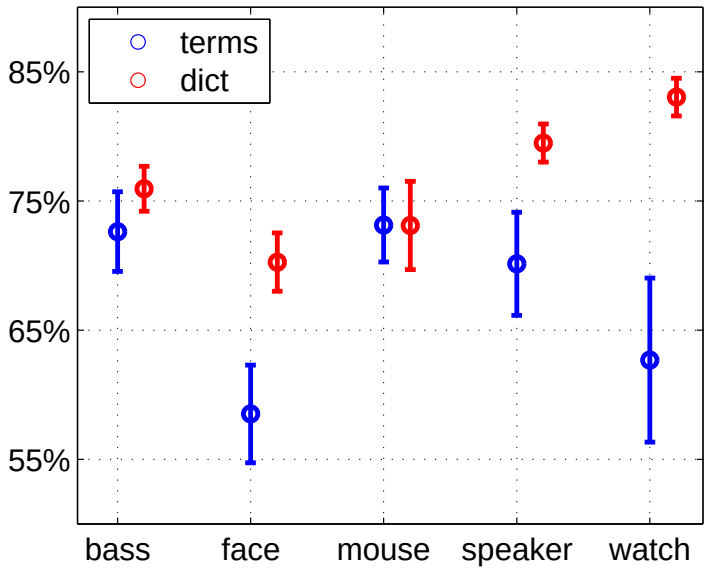


Figure 5-7: Comparison of classification results on the 1-SENSE test set.

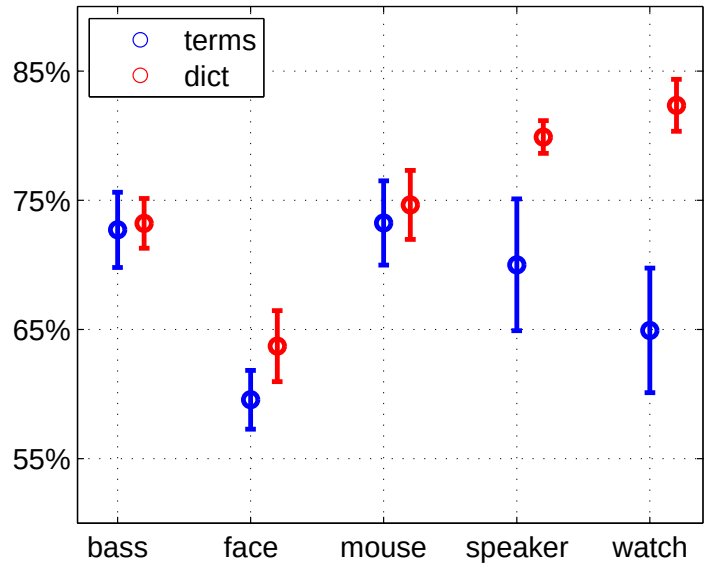


Figure 5-8: Comparison of classification results on the MIX-SENSE test set.

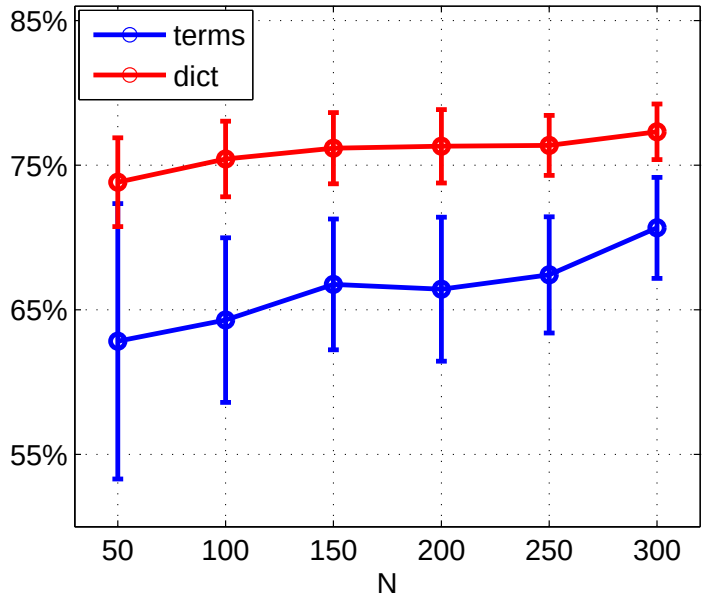


Figure 5-9: Plot of classification results averaged over categories vs. number of training images N on the 1-SENSE test set.

is appropriate for web images. The use of LDA to discover a latent sense space makes the approach robust despite the very limited nature of dictionary definitions. As a final step, a discriminative classifier is trained on the re-ranked mixed-sense images that can predict the correct sense for novel images.

We evaluated our model on a large dataset of web images consisting of search results for several polysemous words. Experiments included retrieval of the ground truth sense and classification of unseen images. On the retrieval task, *WISDOM* improved on the Yahoo search engine precision by re-ranking the images according to sense probability. On the classification task, it outperformed a baseline method that attempts to refine the search by generating sense-specific search terms from Wordnet entries. Classification also improved when the test objects included the other senses of the keyword, and distinctions such as "loudspeaker" vs. "invited speaker" had to be made. To our knowledge, this is the first attempt to automatically deal with polysemy in object recognition.

In this chapter, we assumed that the desired word senses are provided as input to

the algorithm. This can be considered as a form of supervision. This would classify our method as not completely unsupervised, but rather weakly supervised. Also, while this chapter used the WordNet semantic database to obtain sense entries, other repositories could and should be explored. A different source of sense definitions could change not only the senses but also the performance of the model. One avenue for future work is using online encyclopedias, such as Wikipedia, which contain pages rather than sentences of text per entry.

Of course, we would not expect the dictionary senses to always produce accurate visual models, as many senses do not refer to objects (e.g. "bass voice"). Automatic classification of dictionary senses into objects and abstract concepts is a very interesting research question. In the next chapter, we address this question in the framework of the *WISDOM* algorithm, and develop a method that filters senses automatically based on semantic relations in WordNet.

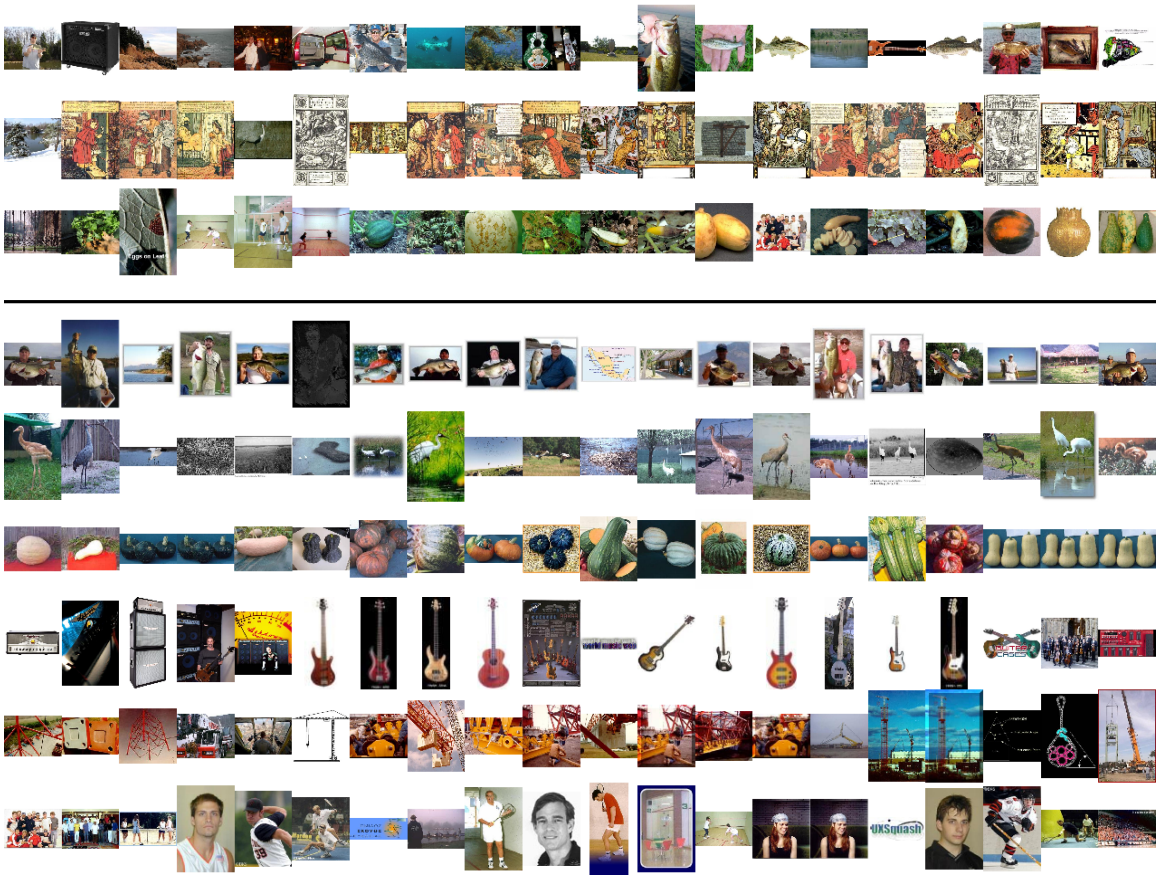


Figure 5-10: The 20 top BASS, CRANE and SQUASH images ranked by Yahoo (top three rows), and by *WISDOM* for each of the ground truthed senses (remaining rows).

6

Automatic Sense Selection

In the previous chapter, we presented *WISDOM*, an unsupervised algorithm for learning visual sense models based on dictionary senses. The algorithm required the sense of the word as an input. In this chapter, we relax the supervision constraints even further, and assume that the target sense of the word is unknown. This brings us even closer to the goal eliminating the need for user input.

6.1 Introduction

The requirement that the dictionary entry corresponding to the desired visual object be specified a priori is a compromise between asking for labeled images and asking for no supervision at all. In many scenarios, the desired senses may be gleaned from the language context. For example, for a collection of image/text documents, a supervised

WSD method can be applied as a pre-processing step to identify the sense of keywords in the vicinity of images. Following that step, visual models can be learned only for the identified set of senses using the *WISDOM* method. In interactive systems that use natural language to interact with the user, the sense can also potentially be inferred from the language context.

However, many practical scenarios call for robots or agents which can learn a visual model on the fly given only a brief spoken or textual definition of an object category. In these scenarios, one cannot always expect to be provided with enough context to identify the correct dictionary sense. A prominent example is the NSF-funded Semantic Vision Robot Challenge (SVRC)[1], which provides robot entrants with a text-file list of categories to be detected shortly before the competition begins. Each participant robot then searches the environment for instances of objects corresponding to the provided terms. More generally, we would like a robot or agent to be able to engage in situated dialog with a human user, and have the robot be able to understand what objects the human is referring to in an environment. While some limited experiments have been carried out on multimodal object recognition [Saenko and Darrell, 2007], it is generally unreasonable to expect that users will limit their vocabulary to existing visual object databases, e.g., Caltech 101/256 or Pascal. We thus would like to take a spoken word from a user's utterance when she is referring to an object of interest, and train a model on the fly so that the robot can find the desired object.

For both the SVRC and the open-vocabulary multimodal object reference problem, and similar tasks, we are therefore faced with the problem of learning a visual model based only on the name of an object. A common approach is to find images on the web that co-occur with the object name by using popular web search services, and train a visual classifier from the search results. As we discussed in previous chapters, words are generally polysemous, and this naive approach can lead to relatively noisy models if many images of clutter senses are added to the model.

Early methods used manual intervention to identify clusters corresponding to the

[1] `http://www.cs.cmu.edu/~srvc/`

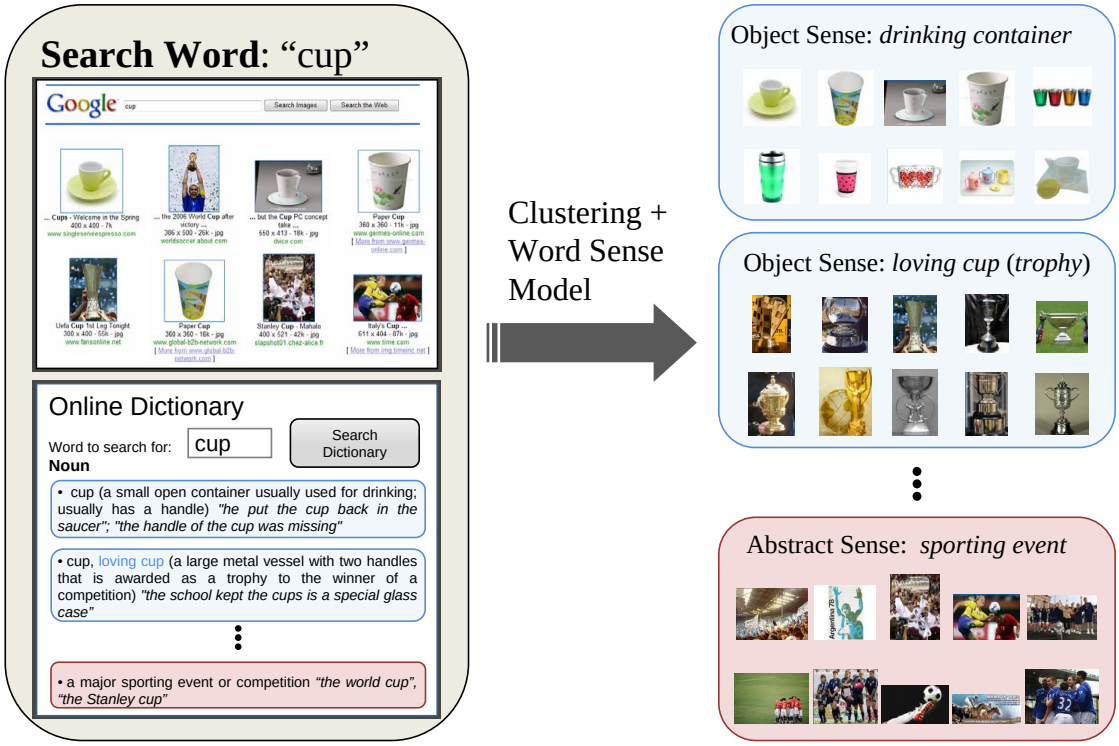


Figure 6-1: Abstract word senses are automatically excluded from the visual model.

desired sense (e.g., Berg and Forsyth [Berg and Forsyth, 2006]). Image clustering methods can group together visually coherent sets of returned queries, but clusters are rarely exactly aligned with actual senses. Individual senses will be split into distinct clusters corresponding to different visual appearances of the object sense, and clutter from senses of the word that are abstract (are not associated with a physical representation; see Figure 6-1) can further complicate matters.

In general, one can imagine using the text and image feature distributions associated with a sense, and/or the words in its dictionary definition, to infer whether a particular sense is one due to a physical entity or a non-physical concept. An alternative approach, one that we explore in this chapter, is to exploit features of the WordNet hierarchy directly to infer whether a sense is abstract or concrete, and thus to form an estimate of the likelihood that a particular image arises from a physical object vs. an abstract concept. Instead of assuming that the physical object senses are

known, we make much more general assumption about the nature of physical objects, namely, that they fall into several general categories of animals, people, artifacts, etc.

We outline our approach to automatic sense selection in Section 6.2. Then, in Section 6.3 we propose an improvement to *WISDOM* that adapts the generic web topics to paired image and text data. In Section 6.4, we show results of detecting concrete senses on three evaluation datasets consisting of web images and their text contexts.

6.2 Selecting Concrete Senses

We tackle the problem of classifying concrete vs. abstract senses in images by extending *WISDOM*, the multimodal sense grounding method presented in the previous chapter. The input is a single word or phrase that maps to a set of senses in WordNet. Given the set of senses, we introduce a step to classify each sense as being abstract or concrete, and consequently either add or remove it from the visual model. We call this model *WISDOM-Concrete*, abbreviated *WISDOM-C*.

We might attempt to accomplish this in a data-driven or supervised way, e.g. by examining the text surrounding each occurrence of the target sense in a sense-tagged corpus to see what role it may play as the subject of a realized action in a sentence. For example, we might learn that the sense of "diamond" that can be the object of the actions of holding, cutting, giving, etc., is an artifact, as opposed to the "rhombus" sense that is an abstract shape.

Fortunately, WordNet contains relatively direct information related to the physicality of a concept. In particular, one of the main functions of WordNet is to put synsets in semantic relation to each other as described in Chapter 5. The semantic network makes it possible to follow the chain of hypernyms all the way to the top of the tree, a node that contains the word "entity". Thus, we can detect a concrete sense by studying its semantic relations to other concepts. For example, we can examine its hypernym to see if it contains synsets such as "artifact", or "animal". What's more, we can restrict the model to include specific types of physical entities: living things,

Table 6.1: WordNet features used in *WISDOM-C*.

Feature	Value
hypernyms	'article', 'instrumentality','article of clothing', 'animal', 'body part'
lexical tag	<artifact>, <animal>, <body>, <plant>

artifacts, clothing, etc.

In addition to semantic relations, WordNet contains lexical file information for each sense in the definition, marking each sense as <state>, <animal>, <person>, <artifact>, etc. For example, for the noun "mouse":

1. <**animal**> mouse *(any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails)*

2. <**state**> shiner, black eye, mouse *(a swollen bruise caused by a blow to the eye)*

3. <**person**> mouse *(person who is quiet or timid)*

4. <**artifact**> mouse, computer mouse *(a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad)*

Table 6.1 summarizes the features that identify a concrete sense. We exclude "people" senses from the present model, as we are not trying to address person (face) recognition. In fact, we remove all *proper noun* senses from the model, excluding people (e.g. Albert Einstein), places (e.g. New York) and other named objects. The reason for doing so is that we are presently concerned with basic object category recognition, and proper nouns refer to specific instances, not categories.

6.3 Topic Adaptation

In the *WISDOM* algorithm presented in the previous chapter, the generative model of sense leverages a latent topic space learned on a large corpus of web pages. These topics generally tend to coincide with different meanings and/or uses of the word. In contrast, text topics occurring in image contexts can be hard to interpret as meanings, and often cluster around specific websites (see discussion in Chapter 4.5). In general, web topics form around both objects and abstract concepts. For example, web topics for MOUSE include computer device topics and topics related to experimentaion on mice in medical research. On the other hand, image context topics form mostly around visual concepts, even though they do not always constitue a coherent object category.

While the generative model has been shown so far to be relatively robust to the presense of abstract web topics, the keywords on which it was tested were highly polysemous, with very distinct meanings. This is not the case for all words. Words that are not strongly polysemous may have several uncommon meanings. As an example, compare the words BASS and STAPLER. BASS has strong distinctions between its common meanings, which show up in both web and image context topics. STAPLER, on the other hand, only has one common meaning and the other topics surrounding the word (e.g. the stapler featured on the comedy show "Office Space") are dictated by the particular document collection.

To better handle the case of mismatched topics, we propose a modification to the *WISDOM* paradigm. The modification involves adapting the web topic to the image context data, in order to better reflect the meanings present in the image collection. Figure 6-2 illustrates this by showing a web topic discovered for FORK (on the left) and the same topic after it has been adapted (right). The original topic seems to be about bicycle forks, however, several words are indicative of the utensil case (these words are enlarged in the figure). After adaptation, most of the words are related to the utensil case, and the topic takes on a decidedly less bicycle-related tone.

Specifically, rather than use the web LDA model directly to model the generation

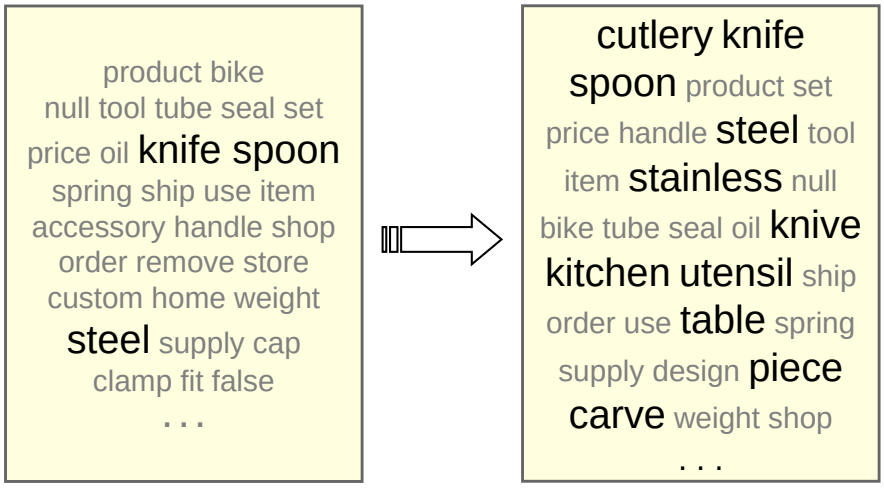


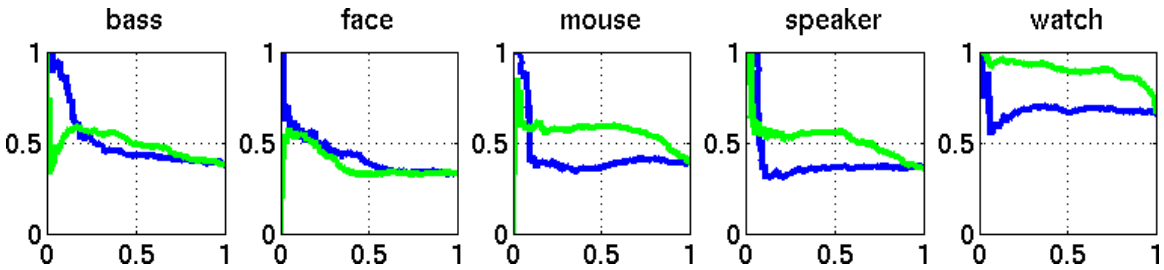
Figure 6-2: A web topic for FORK is adapted to have more likely words related to the utensil sense (shown in large font).

of the text contexts, we employ a semi-supervised paradigm. The topic variables in the LDA model are sampled using Gibbs sampling for several iterations, however, the z variables in the web data are kept fixed. Unlike the sampling procedure used in Chapter 5 to obtain θ parameters for the text contexts, this procedure alters the distributions of words ϕ of the original web topics.

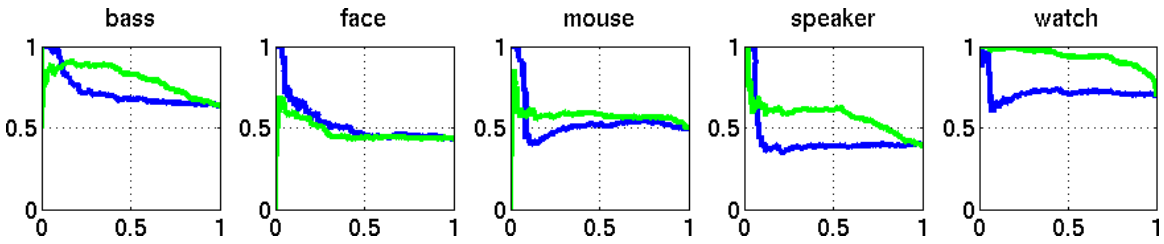
6.4 Experiments

6.4.1 Retrieval of Concrete Senses

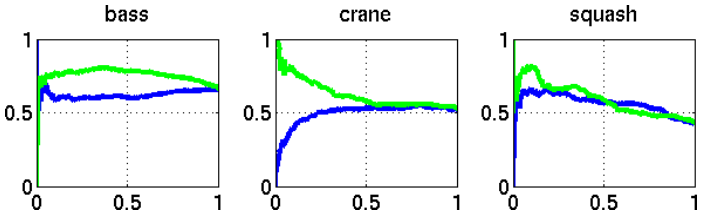
First, we evaluate *WISDOM-C* on the task of retrieving concrete sense images in the MIT-ISD, UIUC-ISD and MIT-OFFICE datasets. Table 6.2 shows the actual concrete senses automatically selected from WordNet entries by our model for each word in the data, using the settings shown in Table 6.1. For the MIT-OFFICE dataset, we restricted the model further to artifact senses and pruned infrequent entries. We also pruned senses that corresponded to parts of objects rather than whole objects, as indicated by the meronym semantic relations. Note that all of the resulting senses shown in Table 6.2 correspond to actual visual concepts and were tagged by human labelers in the datasets.



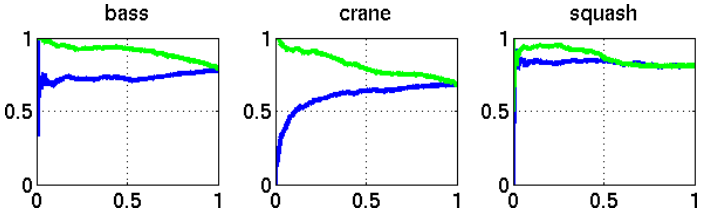
(a) Core, MIT-ISD



(b) Core+Related, UIUC-ISD



(c) Core, UIUC-ISD

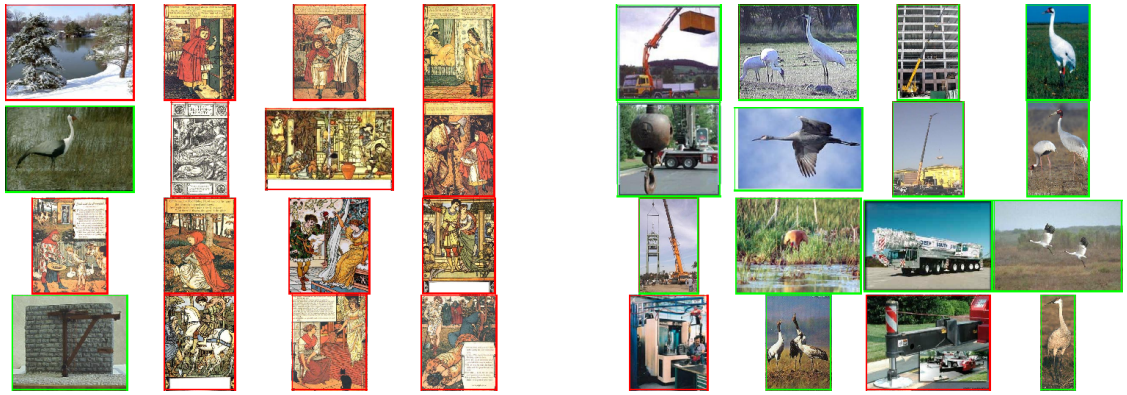


(d) Core+Related, UIUC-ISD

Figure 6-3: Retrieval of concrete senses in MIT-ISD and UIUC-ISD datasets.

MIT Dataset	UIUC-ISD Dataset	OFFICE Dataset
bass-7 (instrument)	bass-7 (instrument)	cellphone-1 (mobile phone)
bass-8 (fish)	bass-8 (fish)	fork-1 (utensil)
face-1 (human face)	crane-4 (machine)	hammer-2 (hand tool)
face-13 (surface)	crane-5 (bird)	keyboard-1 (any keyboard)
mouse-1 (rodent)	squash-1 (plant)	mug-1 (drinking vessel)
mouse-4 (device)	squash-3 (game)	pliers-1 (tool)
speaker-2 (loudspeaker)		scissors-1 (cutting tool)
watch-1 (timepiece)		stapler-1 (stapling device)
		telephone-1 (landline phone)
		watch-1 (timepiece)

Table 6.2: Concrete senses selected from WordNet for words in our datasets.



(a) Yahoo Image Search (b) Concrete-Sense Model

Figure 6-4: The top images returned by the search engine for CRANE, compared to our multimodal concrete-sense model.

In this section, we evaluate the ability of *WISDOM-C* to filter out abstract senses from a given word's image search results. Figure 6-3 shows the resulting recall-precision curves (RPCs), computed by thresholding the probability of any of the concrete senses in a given search result. The ground truth labels used to compute these RPCs are positive if an image was labeled with any *core* sense (Fig.6-3 (a,b)), or any *core* or *related* sense (Fig.6-3 (c,d)) in the dataset, and negative otherwise. These results demonstrate that our model improves the retrieval of images of concrete (visual) senses of words by filtering out the abstract senses. Figure 6-4 shows an example of images being filtered out of the CRANE results, including illustrations by an artist named Crane.

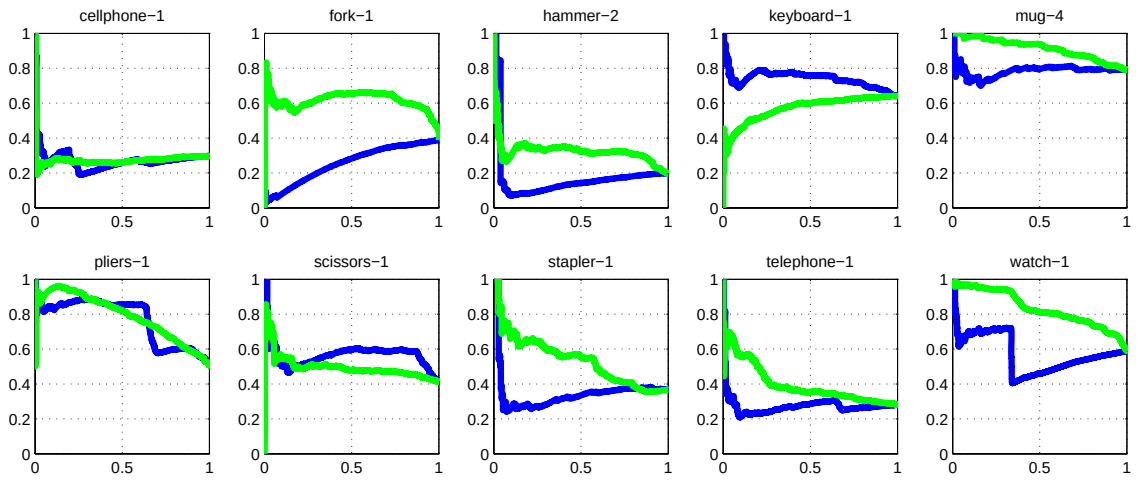


Figure 6-5: Retrieval of concrete senses MIT-OFFICE dataset.

Figure 6-5 shows results on the MIT-OFFICE dataset using the adapted topic version of the model. The average area under the RPC for this data improves from 0.47 for the original search engine order to 0.57 for the adapted-topic *WISDOM* model, and the average RPC area acheived by the non-adapted model is 0.45. Topic adaptation brings a substantial improvement on this data.

In Figure 6-5, the only keyword for which the method causes retrieval to be worse is KEYBOARD. A possible cause for this is that the text topic identified by the model as the most likely to belong to the concrete sense of the word in fact does not describe the canonical keyboard object. A visual inspection of the top results reveals many technical illustrations that have to do with a computer keyboard's use and inner workings, but do not necessarily depict the object in its most recognizable form.

6.4.2 Classification Experiments

We have shown that our method can improve retrieval of concrete senses, therefore providing higher-precision image training data for object recognition algorithms. We have conjectured that this leads to better classification results; in this section, we provide some initial experiments to support this claim. We train multiclass (ten-way)

SVM classifiers using the vocabulary-guided pyramid match kernel over bags of local SIFT features implemented in the LIBPMK library [Lee, 2008]. The training data for the SVM was either the first 100 images returned from the search engine, or the top 100 images ranked by our model. Since we're interested in objects, we keep only <artifact> senses that descend from "instrumentality" or "article". Figure 6-6 shows classification results on held-out test data, averaged over 10 runs on random 80% subsets of the data.

Our method improves accuracy for most of the objects; in particular, classification of "mug" improves greatly due to the non-object senses being filtered out. This is a very difficult task, as evidenced by the baseline performance; the average baseline accuracy is 27%. Training with our method achieves 35% accuracy, a 25% relative improvement. We believe that this relative improvement is due to the higher precision of the training images and will persist even if the overall accuracy were improved due to a better classifier.

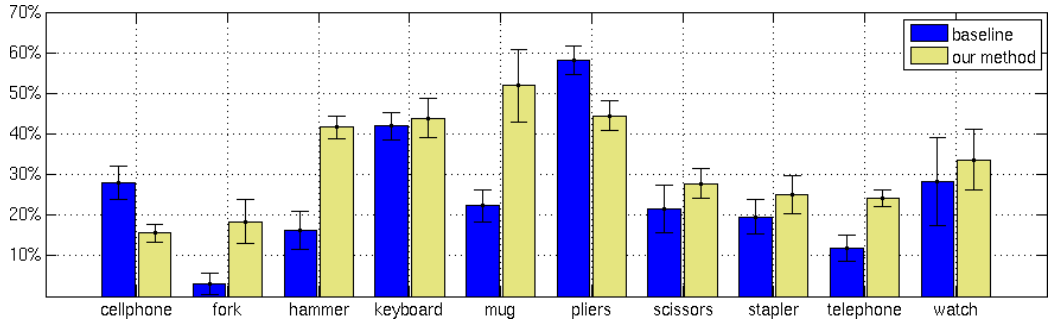


Figure 6-6: Classification accuracy of the ten-way object classifier on MIT-OFFICE.

6.5 Conclusion

We have presented an architecture for clustering image search results for polysemous words based on image and text co-occurrences and grounding latent topics according to dictionary word senses. Our method distinguishes which senses are abstract from those that are concrete, allowing for filtering of the former when constructing a classifier for a particular object of interest to a situated agent. This can be of particular

utility to a mobile robot faced with the task of learning a visual model based only on the name of an object provided on a target list or spoken by a human user.

Our method uses both image features and the text associated with the images to relate estimated latent topics to particular senses in an available online ontology. Our model does not require any human supervision, and takes as input only an English noun. It estimates the probability that a search result is associated with an abstract word sense, rather than a sense that is tied to a physical object. We have carried out experiments with image and text-based models to form estimates of abstract vs. concrete senses, and have shown results detecting concrete-sense images in three multimodal, multi-sense databases. We also demonstrated a 25% relative improvement in accuracy when classifiers are trained with our method as opposed to the raw search results.

7

Multimodal Reference Resolution For Conversational Systems

In this chapter, we present a method that uses both the speech reference and the image to recognize the object identity.[1]

7.1 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 recog-

[1]Portions of this chapter were published in [Saenko and Darrell, 2007]

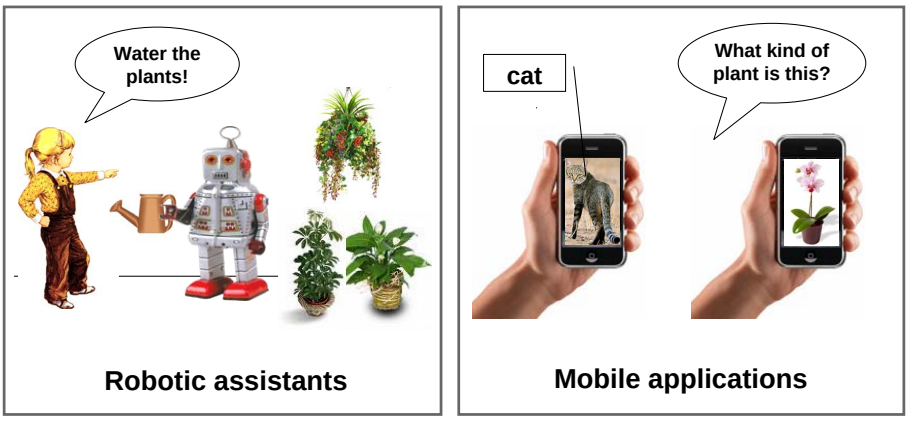


Figure 7-1: Multimodal object reference in conversational systems.

nize 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 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 [Haasch *et al.*, 2005], 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

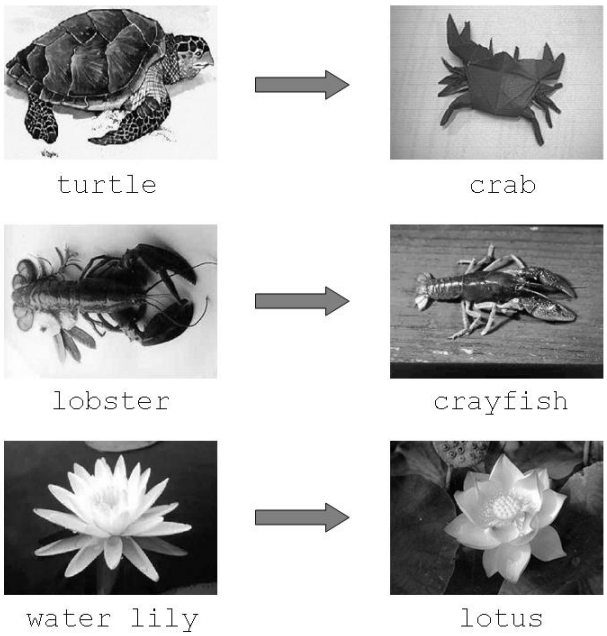


Figure 7-2: Examples of the most visually confusable categories in our dataset (see Section 7.3 for a description of the experiments). The image-based classifier most often misclassified the category on the left as the category on the right.

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 7-2 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. [Potamianos *et al.*, 2003], [Kaiser *et al.*, 2003]). 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.

7.2 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:

$$p(c|x_1, ..., x_m) = \sum_{i=1}^{m} p(c|x_i)\lambda_i, \tag{7.1}$$

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,

$$p(c|x_1, ..., x_m) = \prod_{i=1}^{m} p(c|x_i)^{\lambda_i} \tag{7.2}$$

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 [Chang and Lin, 2001].

7.3 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. The data collection is described in Chapter 3.2. 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.

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%.

7.3.1 Training of Classifiers.

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. The current best-performing object classification methods on *Caltech 101* [Fei-Fei *et al.*, 2007], the image database we use in our experiments, are based on discriminative multi-class classifiers. In [Frome *et al.*, 2006], 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 [Zhang *et al.*, 2006] use a multi-class support vector machine (SVM) classifier with local interest point descriptors as visual features. We use the method of [Grauman and Darrell, 2005], which is also based on a multi-class SVM, but in combination with a kernel that computes distances between pyramids of visual feature histograms.

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 [Grauman and Darrell, 2005], 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 [Grauman and Darrell, 2006], 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 pre-trained 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 7.1, 7.2. 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.

7.3.2 Experimental Settings

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.

7.3.3 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 images per category [Grauman and Darrell, 2007], however, that would not leave enough test images for

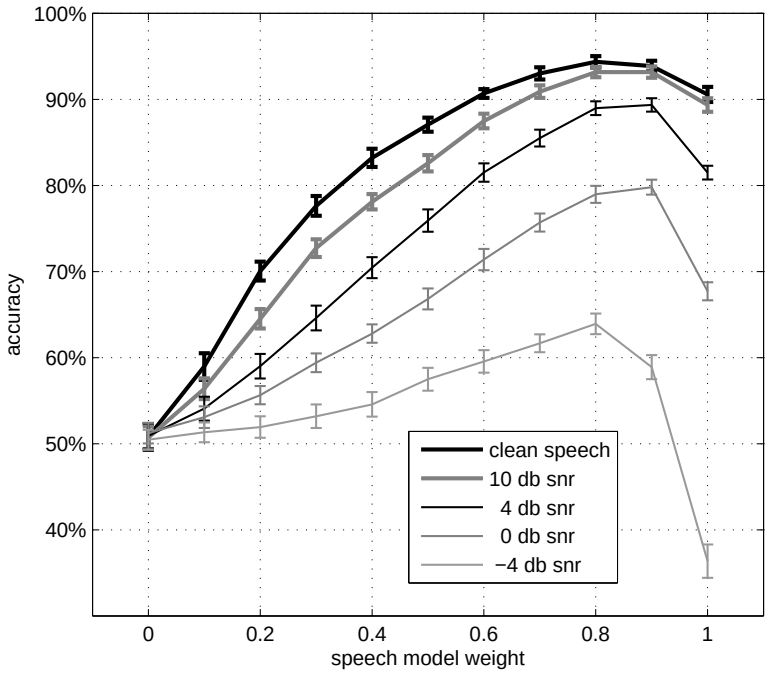


Figure 7-3: Object classification using the mean rule, on the development set. Each line represents the performance on a different level of acoustic noise. The y-axis shows the percent of the samples classified correctly, the x-axis plots the speech weight used for the combined classifier.

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 7-3 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-

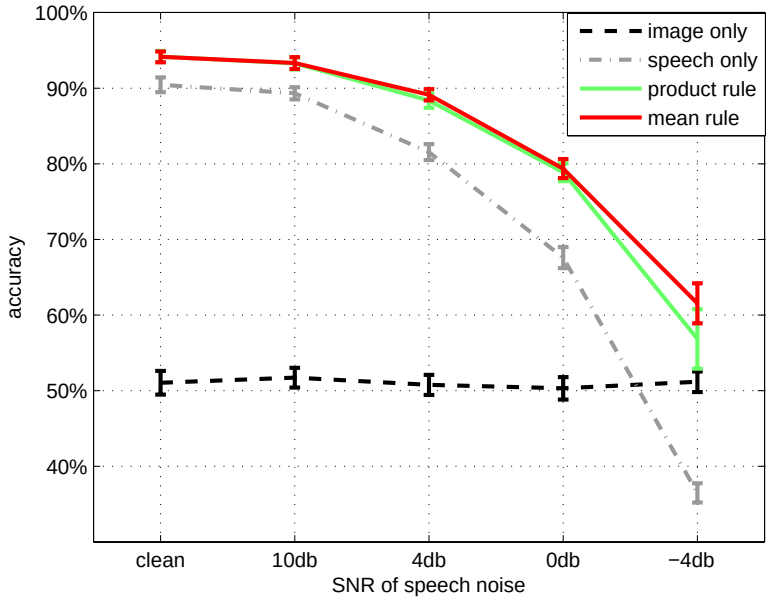


Figure 7-4: Absolute improvement across noise conditions on the test set. The Y-axis shows the percent of the test samples classified correctly, the X-axis shows the SNR of the noise condition. Chance performance is around 1%.

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 7-4. 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.

7.4 Discussion

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 in this chapter 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 incorporating the *WISDOM* approach as a component in the overall multimodal system. 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.

8

Conclusion

Humans interact with their world and with each other in inherently multimodal ways, learning and comminicating about the physical world through the faculties of speech, written language and vision, to name a few. If computers are to match human abilities in this regard, automatic object recognition methods should not be limited to image-space learning. In this dissertation, we have shown that non-traditional information sources, namely, dictionaries, web pages, and spoken utterances, can facilitate object recognition, lessening the need for human supervision and increasing robustness over using image data alone.

Our work shows that massive amounts of parallel image and language data available in electronic form and readily accessible through the Internet can facilitate the automatic acquisition of visual concepts by machine. It advances the state of the art through *WISDOM*, a method for learning visual sense models in the absense of

labeled examples. Human labeling of images is heavily relied-upon in the computer vision community, but it is time-consuming, not online, and must be repeated for every new visual concept. *WISDOM*, inspired by the use of electronic dictionaries to learn word sense models in the natural language community, removes the need for manual labeling of images. The key innovation is the use of the WordNet semantic database together with a collection of web pages to train visual object classifiers. The model is flexible, in that different forms of written knowledge about a visual sense can be used in place of WordNet, such as other dictionaries, encyclopedias or ontologies.

Ours is the first web-based object recognition approach able to predict not only a word label, but also the dictionary meaning of the word. This can be a useful feature at a higher level of interaction, such as speech recognition and discourse processing, as it can make distinctions such as "loudspeaker" vs. "invited speaker". In extensive experiments with both polysemous and single-sense words, we have demonstrated that the version based on WordNet is excellent at retrieving isolated senses from web images. On the task of novel image classification, *WISDOM* outperformed a baseline method that attempts to refine the search by generating sense-specific search terms from Wordnet entries.

Of course, we would not expect all dictionary senses to produce accurate visual models, as many senses do not refer to physical entities. While the question of what constitutes a visual concept remains largely open, this work is a step towards a solution based on the semantic relationships between words. In Chapter 6 we extended *WISDOM* to distinguish abstract senses from those that are more likely to be concrete, allowing it to filter out the abstract ones when constructing a classifier for a particular object. The final model does not require any human supervision, and takes as input only an English noun. For a set of words corresponding to everyday objects, significant improvement in accuracy is obtained when classifiers are trained with our method instead of the unfiltered web search results.

Our unsupervised scheme is of particular utility to an autonomous robot faced with the task of learning a visual model based only on the name of an object, either provided as input or spoken by a human user. In the last part of this dissertation, we

showed that having an a priori visual model of a word can in turn help to disambiguate the user's spoken utterance. Chapter 7 presented a multimodal object category classifier that combines image-only and speech-only hypotheses in a probabilistic way, and demonstrated that combining the modalities improves recognition across several audio noise levels. We have thus shown that it may be advantageous for HRI systems to use both channels to recognize object references, when possible, as opposed to the conventional approach of relying on speech-only or image-only recognition.

8.1 Limitations and Future Work

At the end of each method chapter, we have summarized any outstanding technical issues and future work directions pertaining to that specific component of our system. Here we discuss the "big picture" view of what is still missing and where this line of research might lead us next.

An Open Vocabulary of Concepts. Ironically, our experiments with a system that learns visual concepts in an unsupervised way were limited by the lack of labeled images to test it on. Although the datasets used for evaluation contained a total of over 44,000 images, there were only 17 unique words tested. An important part of continuing this research is to test the ideas on a dataset of labeled web images of much grander scale. Fortunately, such a dataset may soon be available in the form of ImageNet [Deng *et al.*, 2009]. Also, a very interesting research direction is the question of visual vs. abstract concepts. Can we determine automatically if a word or phrase in a passage of text refers to an physical object or an abstract idea? This may be more difficult than it seems at first thought, as abstract concepts are frequently used by manufacturers to brand products.

Adaptive HCI System. Chapter 7 proved that image and voice provide complimentary cues of object identity, however, it was limited to a small vocabulary. The limiting factor was the lack of image-based object classifiers for arbitrary words. The next step is to use *WISDOM* to expand the vocabulary of the overall system. The final system could then recognize user references to arbitrary objects. Another future

work direction is to adapt the models built from web images to be useful for an autonomous robot to understand its environment. Web images are not representative of the types of images that a robot would come across in an office or home environment. Images on the web are typically taken by professional photographers and aimed for an aestheticaly pleasing effect. As a result, these images have little blurring or occlusion, and the objects are often centered and in canonical poses. On the other hand, a robot in the real world would encounter images with poor lighting, blurring and random poses. Our method could be used to robustly process user references to an object in a home tour scenario, providing labeled examples for adaptation of the prior model of the object.

A

Word Definitions

This appendix includes the WordNet definitions of words used as queries to collect the datasets described in Chapter 3.

Table A.1: WordNet definitions for words in the datasets.

Synset	Definition
BASS-1	the lowest part of the musical range
BASS-2, BASS-PART-1	the lowest part in polyphonic music
BASS-3, BASSO-1	an adult male singer with the lowest voice
Continued on next page	

Table A.1 – continued from previous page

Synset	Definition
SEA BASS-1, BASS-4	the lean flesh of a saltwater fish of the family Serranidae
FRESHWATER-BASS-1, BASS-5	any of various North American freshwater fish with lean flesh and especially of the genus Micropterus
BASS-6, BASS-VOICE-1, BASSO-2	the lowest adult male singing voice
BASS-7	the member with the lowest range of a family of musical instruments
BASS-8	nontechnical name for any of numerous edible marine and freshwater spiny-finned fishes
CELLULAR TELEPHONE-1, CELLULAR PHONE-1, CELLPHONE-1, CELL-5, MOBILE PHONE-1	a hand-held mobile radiotelephone for use in an area divided into small sections, each with its own short-range transmitter/receiver
CRANE-1, STEPHEN CRANE-1	United States writer (1871-1900)
CRANE-2, HART CRANE-1, HAROLD HART CRANE-1	United States poet (1899-1932)
GRUS-1, CRANE-3	a small constellation in the southern hemisphere near Phoenix
CRANE-4	lifts and moves heavy objects; lifting tackle is suspended from a pivoted boom that rotates around a vertical axis

Continued on next page

Table A.1 – continued from previous page

Synset	Definition
CRANE-5	large long-necked wading bird of marshes and plains in many parts of the world
FACE-1, HUMAN FACE-1	the front of the human head from the forehead to the chin and ear to ear) "he washed his face"; "I wish I had seen the look on his face when he got the news"
EXPRESSION-1, LOOK-1, ASPECT-5, FACIAL EXPRESSION-2, FACE-2	the feelings expressed on a person's face) "a sad expression"; "a look of triumph"; "an angry face"
FACE-3	the general outward appearance of something) "the face of the city is changing"
FACE-4	the striking or working surface of an implement
FACE-5	a part of a person that is used to refer to a person) "he looked out at a roomful of faces"; "when he returned to work he met many new faces"
SIDE-4, FACE-6	a surface forming part of the outside of an object) "he examined all sides of the crystal"; "dew dripped from the face of the leaf"
FACE-7	the part of an animal corresponding to the human face
FACE-8	the side upon which the use of a thing depends (usually the most prominent surface of an object)) "he dealt the cards face down"
GRIMACE-1, FACE-9	a contorted facial expression) "she made a grimace at the prospect"

Continued on next page

Table A.1 – continued from previous page

Synset	Definition
FONT-1, FOUNT-1, TYPEFACE-1, FACE-10, CASE-14	a specific size and style of type within a type family
FACE-11	status in the eyes of others) "he lost face"
BOLDNESS-2, NERVE-3, BRASS-4, FACE-12, CHEEK-4	impudent aggressiveness) "I couldn't believe her boldness"; "he had the effrontery to question my honesty"
FACE-13	a vertical surface of a building or cliff
FORK-1	cutlery used for serving and eating food
BRANCHING-1, RAMIFICATION-1, FORK-2, FORKING-2	the act of branching out or dividing into branches
FORK-3, CROTCH-1	the region of the angle formed by the junction of two branches) "they took the south fork"; "he climbed into the crotch of a tree"
FORK-4	an agricultural tool used for lifting or digging; has a handle and metal prongs
CROTCH-2, FORK-5	the angle formed by the inner sides of the legs where they join the human trunk
HAMMER-1, COCK-3	the part of a gunlock that strikes the percussion cap when the trigger is pulled

Continued on next page

Table A.1 – continued from previous page

Synset	Definition
HAMMER-2	a hand tool with a heavy rigid head and a handle; used to deliver an impulsive force by striking
MALLEUS-1, HAMMER-3	the ossicle attached to the eardrum
MALLET-2, HAMMER-4	a light drumstick with a rounded head that is used to strike such percussion instruments as chimes, kettledrums, marimbas, glockenspiels, etc.
HAMMER-5	a heavy metal sphere attached to a flexible wire; used in the hammer throw
HAMMER-6	a striker that is covered in felt and that causes the piano strings to vibrate
HAMMER-7, POWER HAMMER-1	a power tool for drilling rocks
HAMMER-8, POUND-14, HAMMERING-1, POUNDING-3	the act of pounding (delivering repeated heavy blows)) "the sudden hammer of fists caught him off guard"; "the pounding of feet on the hallway"
KEYBOARD-1	device consisting of a set of keys on a piano or organ or typewriter or typesetting machine or computer or the like
KEYBOARD-2	holder consisting of an arrangement of hooks on which keys or locks can be hung
MOUSE-1	any of numerous small rodents typically resembling diminutive rats having pointed snouts and small ears on elongated bodies with slender usually hairless tails

Continued on next page

Table A.1 – continued from previous page

Synset	Definition
SHINER-1, BLACK EYE-1, MOUSE-2	a swollen bruise caused by a blow to the eye
MOUSE-3	person who is quiet or timid
MOUSE-4, COMPUTER MOUSE-1	a hand-operated electronic device that controls the coordinates of a cursor on your computer screen as you move it around on a pad; on the bottom of the device is a ball that rolls on the surface of the pad) "a mouse takes much more room than a trackball"
MUG-1, MUGFUL-1	the quantity that can be held in a mug
CHUMP-1, FOOL-2, GULL-1, MARK-9, PATSY-1, FALL GUY-1, SUCKER-1, SOFT TOUCH-1, MUG-2	a person who is gullible and easy to take advantage of

Continued on next page

Table A.1 – continued from previous page

Synset	Definition
COUNTENANCE-3, PHYSIOGNOMY-1, PHIZ-2, VISAGE-1, KISSER-2, SMILER-2, MUG-3	the human face ('kisser' and 'smiler' and 'mug' are informal terms for 'face' and 'phiz' is British)
MUG-4	with handle and usually cylindrical
PLIER-1, PLYER-1	someone who plies a trade
PLIERS-1, PAIR OF PLIERS-1, PLYERS-1	a gripping hand tool with two hinged arms and (usually) serrated jaws
SCISSORS-1, PAIR OF SCISSORS-1	an edge tool having two crossed pivoting blades
SCISSORS-2, SCISSORS HOLD-1, SCISSOR HOLD-1, SCISSOR GRIP-1, SCISSORS GRIP-1	a wrestling hold in which you wrap your legs around the opponents body or head and put your feet together and squeeze
SCISSORS-3	a gymnastic exercise performed on the pommel horse when the gymnast moves his legs as the blades of scissors move

Continued on next page

Table A.1 – continued from previous page

Synset	Definition
SPEAKER-1, TALKER-1, UTTERER-3, VERBALIZER-1, VERBALISER-1	someone who expresses in language; someone who talks (especially someone who delivers a public speech or someone especially garrulous)) "the speaker at commencement"; "an utterer of useful maxims"
LOUDSPEAKER-1, SPEAKER-2, SPEAKER UNIT-1, LOUDSPEAKER SYSTEM-1, SPEAKER SYSTEM-1	electro-acoustic transducer that converts electrical signals into sounds loud enough to be heard at a distance
SPEAKER-3	the presiding officer of a deliberative assembly) "the leader of the majority party is the Speaker of the House of Representatives"
SQUASH-1, SQUASH VINE-1	any of numerous annual trailing plants of the genus Cucurbita grown for their fleshy edible fruits
SQUASH-2	edible fruit of a squash plant; eaten as a vegetable
SQUASH-3, SQUASH RACQUETS-1, SQUASH RACKETS-1	a game played in an enclosed court by two or four players who strike the ball with long-handled rackets
STAPLER-1, STAPLING MACHINE-1	a machine that inserts staples into sheets of paper in order to fasten them together
TELEPHONE-1, PHONE-1, TELEPHONE SET-1	electronic equipment that converts sound into electrical signals that can be transmitted over distances and then converts received signals back into sounds) "I talked to him on the telephone"

Continued on next page

Table A.1 – continued from previous page

Synset	Definition
TELEPHONE-2, TELEPHONY-1	transmitting speech at a distance
WATCH-1, TICKER-2	a small portable timepiece
WATCH-2	a period of time (4 or 2 hours) during which some of a ship's crew are on duty
WATCH-3, VIGIL-3	a purposeful surveillance to guard or observe
WATCH-4	the period during which someone (especially a guard) is on duty
LOOKOUT-1, LOOKOUT MAN-1, SENTINEL-1, SENTRY-1, WATCH-5, SPOTTER-3, SCOUT-1, PICKET-1	a person employed to keep watch for some anticipated event
VIGIL-2, WATCH-6	the rite of staying awake for devotional purposes (especially on the eve of a religious festival)

Bibliography

[Agirre and Edmonds, 2006] Eneko Agirre and Philip Edmonds. *Word Sense Disambiguation: Algorithms and Applications (Text, Speech and Language Technology).* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[Barnard and Johnson, 2005] Kobus Barnard and Matthew Johnson. Word sense disambiguation with pictures. *Artif. Intell.*, 167(1-2):13–30, 2005.

[Barnard *et al.*, 2003] Kobus Barnard, Pinar Duygulu, David A. Forsyth, Nando de Freitas, David M. Blei, and Michael I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[Berg and Forsyth, 2006] Tamara L. Berg and David A. Forsyth. Animals on the web. In *CVPR*, pages 1463–1470, 2006.

[Biederman, 1987] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychol Rev*, 94(2):115–147, Apr 1987.

[Bilmes and Kirchhoff, 2000] Jeff Bilmes and Katrin Kirchhoff. Directed graphical models of classifier combination: Application to phone recognition. In *ICSLP*, Beijing, China, October 2000.

[Blei and Jordan, 2003] David M. Blei and Michael I. Jordan. Modeling annotated data. In *SIGIR '03: Proc. 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 127–134, New York, NY, USA, 2003. ACM.

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

[Blei, 2004] David M. Blei. *Probabilistic Models of Text and Images.* PhD thesis, U.C. Berkeley, Division of Computer Science, 2004.

[Blum and Mitchell, 1998] Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *COLT' 98: Proc. 11th annual conference on Computational learning theory*, pages 92–100, New York, NY, USA, 1998. ACM.

[Bolt, 1980] Richard A. Bolt. "put-that-there": Voice and gesture at the graphics interface. In *SIGGRAPH '80: Proc. 7th annual conference on Computer graphics and interactive techniques*, pages 262–270, New York, NY, USA, 1980. ACM.

[Cai *et al.*, 2004] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, and Ji-Rong Wen. Hierarchical clustering of www image search results using visual, textual and link information. In *MULTIMEDIA '04: Proc. 12th annual ACM international conference on Multimedia*, pages 952–959, New York, NY, USA, 2004. ACM.

[Chang and Lin, 2001] Chih-Chung Chang and Chih-Jen Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Collins *et al.*, 2008] Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei. Towards scalable dataset construction: An active learning approach. In *ECCV '08: Proc. 10th European Conference on Computer Vision*, pages 86–98, Berlin, Heidelberg, 2008. Springer-Verlag.

[Corel, 2009] Corel. *Corel stock photo images*, 2009. `http://www.corel.com`.

[Deerwester *et al.*, 1990] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407, 1990.

[Deng *et al.*, 2009] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009.

[Everingham *et al.*,] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf.

[Fei-Fei *et al.*, 2007] Li Fei-Fei, Robert Fergus, and Pietro Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[Feng *et al.*, 2004] Huamin Feng, Rui Shi, and Tat-Seng Chua. A bootstrapping framework for annotating and retrieving www images. In *MULTIMEDIA '04: Proc. 12th annual ACM international conference on Multimedia*, pages 960–967, New York, NY, USA, 2004. ACM.

[Fergus *et al.*, 2005] Robert Fergus, Li Fei-Fei, Pietro Perona, and Andrew Zisserman. Learning object categories from google's image search. In *ICCV*, pages 1816–1823, 2005.

[Fink and Ullman, 2008] Michael Fink and Shimon Ullman. From aardvark to zorro: A benchmark for mammal image classification. *Int. J. Comput. Vision*, 77(1-3):143–156, 2008.

[Frome *et al.*, 2006] Andrea Frome, Yoram Singer, and Jitendra Malik. Image retrieval and classification using local distance functions. In *NIPS*, pages 417–424, 2006.

[Google, 2009] Google. *Google Similar Images^{TM}*, 2009. `http://similar-images.googlelabs.com`.

[Grauman and Darrell, 2005] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, pages 1458–1465, 2005.

[Grauman and Darrell, 2006] Kristen Grauman and Trevor Darrell. Approximate correspondences in high dimensions. In *NIPS*, pages 505–512, 2006.

[Grauman and Darrell, 2007] Kristen Grauman and Trevor Darrell. The pyramid match kernel: Efficient learning with sets of features. *J. Mach. Learn. Res.*, 8:725–760, 2007.

[Griffin *et al.*, 2007] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, 2007.

[Griffiths and Steyvers, 2004] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proc. National Academy of Sciences*, 101:5228–5235, 2004.

[Griffiths *et al.*, 2004] Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *NIPS*, 2004.

[Haasch *et al.*, 2005] A. Haasch, N. Hofemann, J. Fritsch, and G. Sagerer. A multi-modal object attention system for a mobile robot. In *Proc. IEEE/RSJ Int. Conf. on Intelligent Robots and Systems*, pages 1499–1504, Edmonton, Alberta, Canada, August 2005. IEEE.

[Hofmann, 1999] Thomas Hofmann. Probabilistic latent semantic indexing. In *SIGIR '99: Proc. 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA, 1999. ACM.

[Hsu and Glass, 2006] Bo-June (Paul) Hsu and James Glass. Style & topic language model adaptation using hmm-lda. In *Empirical Methods in Natural Language Processing (EMNLP)*, Sydney, Australia, 2006.

[Jain *et al.*, 2007] Vidit Jain, Erik G. Learned-Miller, and Andrew McCallum. People-lda: Anchoring topics to people using face recognition. In *ICCV*, pages 1–8, 2007.

[James *et al.*, 2003] Yixin Chen James, James Z. Wang, and Robert Krovetz. An unsupervised learning approach to content-based image retrieval. In *IEEE Proc. Inter. Symposium on Signal Processing and Its Applications*, pages 197–200, 2003.

[Kaiser *et al.*, 2003] Edward C. Kaiser, Alex Olwal, David McGee, Hrvoje Benko, Andrea Corradini, Xiaoguang Li, Philip R. Cohen, and Steven Feiner. Mutual disambiguation of 3d multimodal interaction in augmented and virtual reality. In *ICMI*, pages 12–19, 2003.

[Kittler *et al.*, 1998] Josef Kittler, Mohamad Hatef, Robert P. W. Duin, and Jiri Matas. On combining classifiers. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(3):226–239, 1998.

[Larlus and Jurie, 2009] Diane Larlus and Frédéric Jurie. Latent mixture vocabularies for object categorization and segmentation. *Image Vision Comput.*, 27(5):523–534, 2009.

[Lee, 2008] John J. Lee. Libpmk: A pyramid match toolkit. Technical Report MIT-CSAIL-TR-2008-17, MIT Computer Science and Artificial Intelligence Laboratory, April 2008. Software available at: `http://people.csail.mit.edu/jjl/libpmk/`.

[Li *et al.*, 2007] Li-Jia Li, Gang Wang, and Li Fei-Fei. Optimol: automatic online picture collection via incremental model learning. In *CVPR*, 2007.

[Liu *et al.*, 2007] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recogn.*, 40(1):262–282, 2007.

[Loeff *et al.*, 2006] Nicolas Loeff, Cecilia Ovesdotter Alm, and David A. Forsyth. Discriminating image senses by clustering with multimodal features. In *ACL*, 2006.

[Mikolajczyk and Schmid, 2004] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.

[Navigli, 2009] Roberto Navigli. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):1–69, 2009.

[Nigam *et al.*, 2000] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. In *Machine Learning*, pages 103–134, 2000.

[Ponce *et al.*, 2006] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman. Dataset issues in object recognition. In *Toward Category-Level Object Recognition, volume 4170 of LNCS*, pages 29–48. Springer, 2006.

[Porter, 1988] M. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1988.

[Potamianos *et al.*, 2003] Gerasimos Potamianos, Chalapathy Neti, Guillaume Gravier, Ashutosh Garg, and Andrew W. Senior. Recent advances in the automatic recognition of audiovisual speech. In *Proc. IEEE*, pages 1306–1326, 2003.

[Rosen-Zvi *et al.*, 2004] Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *UAI*, pages 487–494, 2004.

[Roy *et al.*, 2002] Deb Roy, Peter Gorniak, Niloy Mukherjee, and Josh Juster. A trainable spoken language understanding system for visual object selection. In *Proc. International Conference of Spoken Language Processing*, 2002.

[Russell *et al.*, 2008] Bryan C. Russell, Antonio B. Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *International Journal of Computer Vision*, 77(1-3):157–173, 2008.

[Saenko and Darrell, 2007] Kate Saenko and Trevor Darrell. Object category recognition using probabilistic fusion of speech and image classifiers. In *MLMI*, pages 36–47, 2007.

[Saenko and Darrell, 2008] Kate Saenko and Trevor Darrell. Unsupervised learning of visual sense models for polysemous words. In *NIPS*, pages 1393–1400, 2008.

[Saenko *et al.*, 2005] Kate Saenko, Karen Livescu, Michael Siracusa, Kevin Wilson, James R. Glass, and Trevor Darrell. Visual speech recognition with loosely synchronized feature streams. In *ICCV*, pages 1424–1431, 2005.

[Saenko *et al.*, 2009] Kate Saenko, Karen Livescu, James Glass, and Trevor Darrell. Multistream articulatory feature-based models for visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(9):1700–1707, 2009.

[Schroff *et al.*, 2007] Florian Schroff, Antonio Criminisi, and Andrew Zisserman. Harvesting image databases from the web. In *ICCV*, pages 1–8, 2007.

[Steyvers and Griffiths,] M. Steyvers and T. Griffiths. Matlab topic modeling toolbox. Available at `http://psiexp.ss.uci.edu/research/programs\_data/toolbox.htm`.

[Teh *et al.*, 2003] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101, 2003.

[Torralba *et al.*, 2008] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(11):1958–1970, 2008.

[Wei and Croft, 2006] Xing Wei and Bruce W. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proc. 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, USA, 2006. ACM Press.

[Wikipedia, 2009] *Wikipedia, a free online encyclopedia*, 2009. `http://en.wikipedia.org`.

[Yao *et al.*, 2007] B. Yao, X. Yang, and S.C. Zhu. Introduction to a large-scale general purpose ground truth database: Methodology, annotation tool and benchmarks. In *EMMCVPR*, pages 169–183, 2007.

[Yarowsky, 1995] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *ACL*, pages 189–196, 1995.

[Yu *et al.*, 2008] Shipeng Yu, Balaji Krishnapuram, Romer Rosales, Harald Steck, and Bharat R. Rao. Bayesian co-training. In *NIPS*, pages 1665–1672. MIT Press, Cambridge, MA, 2008.

[Zhang *et al.*, 2006] Hao Zhang, Alexander C. Berg, Michael Maire, and Jitendra Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, pages 2126–2136, 2006.