Semi-Supervised Learning of Visual Classifiers from Web Images and Text

Nicholas Morsillo[1], **Christopher Pal**[1,2], **Randal Nelson**[1]

{morsillo,cpal,nelson}@cs.rochester.edu

[1]Department of Computer Science
University of Rochester
Rochester, NY, USA

[2]Département de génie informatique et génie logiciel
École Polytechnique de Montréal
Montréal, QC, Canada

Abstract

The web holds tremendous potential as a source of training data for visual classification. However, web images must be correctly indexed and labeled before this potential can be realized. Accordingly, there has been considerable recent interest in collecting imagery from the web using image search engines to build databases for object and scene recognition research. While search engines can provide rough sets of image data, results are noisy and this leads to problems when training classifiers. In this paper we propose a semi-supervised model for automatically collecting clean example imagery from the web. Our approach includes both visual and textual web data in a unified framework. Minimal supervision is enabled by the selective use of generative and discriminative elements in a probabilistic model and a novel learning algorithm. We show through experiments that our model discovers good training images from the web with minimal manual work. Classifiers trained using our method significantly outperform analogous baseline approaches on the Caltech-256 dataset.

1 Introduction

With the advent of the internet we are accustomed to having practically any information we want within moments of a search query. Search engines have become strikingly accurate at delivering relevant text-based information from the web, but the same cannot be said for web image content. There are billions of images on the web yet most of them are not indexed effectively. Web image search tools return copious results, but these are contaminated with a high degree of label noise. If we can filter relevant images from noisy search results, the web becomes an attractive source of training imagery for visual classifiers.

There are a handful of general approaches for specifying and refining web-based visual queries. Traditionally, queries take the form of text keywords and images are indexed by their textual metadata. The drawback to this approach is found in the lack of quality metadata for images on the web. An alternative approach is to formulate an image based query and retrieve images by visual similarity, but this is challenging given the vast amount of image data on the web.

In this paper we present a hybrid method to capture the benefits of both textual and visual queries. One can use existing image search technology to provide a rough set of images from which a user selects a small number of examples depicting the visual concept of interest.

Once a query is captured in the form of one or a few example images, the task becomes a matter of filtering noisy web image search data to discover visually related images. If the example images are selected from the web search results, they are associated with html documents which provide additional textual information about the query. We can therefore construct a model which combines text and image features to distinguish good images among noisy web data. Our method allows one to harness the power of existing text-based image search engines while performing a deeper visual analysis on a small set of roughly labeled images.

The model we present is unique for the following reasons. It is a tailored probabilistic graphical model containing both directed and undirected components in order to combine different feature types effectively. We explicitly handle the difficult task of learning from minimal training data (a small set of query images) by developing a semi-supervised technique based on a novel hybrid expectation maximization / expected gradient [Salakhutdinov *et al.*, 2003; Dempster *et al.*, 1977] procedure. Our approach has the added benefit that it is fast and fits nicely within the framework of existing image search technologies.

We show in our experiments that images returned by our model are significantly more accurate than traditional image search with respect to a user-defined visual query. We find that both text and image features are useful, sometimes to different degrees depending on the query words. Additionally, we show that classifiers trained with web data using our semi-supervised approach outperform analogous classifiers on standard testing data sets. The methods presented can be used to easily collect more training data to enhance existing state of the art visual classification algorithms.

2 Related Work

The prospect of learning visual classifiers from web data has started to receive considerable research attention [Ponce *et al.*, 2006; Torralba *et al.*, 2007]. For example, [Ponce *et al.*,

2006] emphasize the general need for better object category datasets. Our approach here as well as the others we briefly review represent parts of the solution to this problem. In the following review we emphasize recent work relevant to the problem of learning visual classifiers from *noisy* search results. This problem can be viewed as synonymous with the goal of extracting a clean set of example images for the purposes of dataset construction.

[Fergus *et al.*, 2005] examine the problem of learning object categories from Google data. A variant of the pLSA clustering method is developed to successfully learn category models from noisy web data. An interesting component to this work is the method of selecting training data. The top few image search results are chosen as the training set since these results tend to be more accurate. To collect even more potential training images each query is submitted in multiple languages. [Li-Jia Li and Fei-Fei, 2007] also take an image-only approach. A hierarchical Dirichlet process model is used to retrieve clean image sets for a number of categories from web image search. [Schroff *et al.*, 2007] is another recent work which successfully filters web image sets using image content alone.

[Berg and Forsyth, 2006] look at a combination of web text and image features for the creation of a themed dataset consisting of types of animals. A pre-clustering is performed on the text via LDA, and supervision enters the process by the manual choice of relevant LDA clusters. We find via experiment that this procedure is ineffective for most object categories due to a lack of immediately clear text cluster and visual group correspondence.

Numerous other works have considered visual concept learning in the context of content based image retrieval (CBIR). Notably, [Rui *et al.*, 1998] include user feedback in the query to account for subjectivity of relevance results. [Lu *et al.*, 2000] include both low level visual features and high level semantics for improved query accuracy.

At the same time as this increasing interest in leveraging the web for dataset construction, there has been an increasing focus in the Machine Learning and Data Mining communities on the tradeoff between generative and discriminative approaches to pattern recognition [Ng and Jordan, 2001; McCallum *et al.*, 2006; Lasserre *et al.*, 2006]. A generative probabilistic model attempts to capture the joint probability distribution of data points x and labels y, whereas a discriminative approach models only the conditional $p(y|x)$. Under the correct model assumptions and data conditions one approach can perform significantly better than the other. Typically, generative models tend to succeed when the model makes fair independence assumptions about the data and the data is sparse. Discriminative approaches are often better when the underlying data distribution $p(x)$ is difficult to model, or when the dataset size grows toward infinity. Our technique applies each of these approaches on portions of the data where they are expected to perform best using a probability model with both discriminative and generative components. This permits effective training of discriminative components in a weakly-supervised setting. Our model is also principled in the sense that it makes no ad-hoc modifications to a likelihood objective defined by an underlying probabilistic model.

3 A Model for Web Images and Text

In the following section we describe our proposed model. We begin with an overview of our feature extraction methods for images and web text. We then explain the model structure and methods for parameter estimation in supervised and semi-supervised scenarios.

3.1 Image and Text Features

Histograms of quantized SIFT descriptors [Lowe, 1999] are used to represent image data in our model. SIFT image descriptors are vectors composed of spatially arranged histograms of image gradient information. Descriptors are sampled at each point on a closely spaced grid over the image at multiple scales. Each descriptor is aligned with the dominant gradient direction in the image region. We forego the common approach of interest point detection in light of recent evidence that it is not necessarily more effective than dense grid sampling [Nowak *et al.*, 2006]. The descriptors are enhanced with local color information by concatenation with an 11-dimensional color histogram vector. The color histogram belonging to each SIFT feature is computed in the local image area surrounding the feature and binned along 11 principal color axes. We find experimentally that the addition of color information to SIFT vectors improves performance for recognition tasks.

We pre-compute a visual codebook by application of k-means to a large pool of randomly sampled descriptors. Each descriptor extracted from an image is assigned to its nearest word in the codebook, and an image is represented as a vector of visual word counts. The visual codebook size is fixed to 400 for experiments involving the musical instruments dataset of section 4.1 in order to match common practices in the literature. An optimal codebook size of 10,000 was determined experimentally for work involving the Caltech 256 dataset in section 4.4.

Text inputs to our model are derived from associated html documents via a latent topic model. We use the Latent Dirichlet Allocation document-topic model [Blei *et al.*, 2003] to learn a robust low-dimensional representation. LDA is a generative model of documents where each document is a multinomial distribution over a set of unobserved topics, and each topic is a multinomial over words. Text processing for web images begins with the 100 words closest to the html image anchor. The words are filtered by a standard stoplist which is augmented to remove common web-related junk words. We use the LDA implementation from the Mallet Toolkit [McCallum, 2002] to train a model with 10 topics on the set of word vectors for a query. Then, each document's text feature is given by the 10-vector of mixing proportions for the 10 topics. The number of topics was optimized empirically by looking for meaningful semantic clusterings in the topic-word assignments.

The application of LDA as a pre-processing step might seem unusual. We note that principal component analysis

(PCA) is commonly applied to real valued data to both reduce the dimensionality of input data and to reduce correlations among the dimensions of the transformed data. As such, simpler models with feature independence assumptions (such as diagonal covariance Gaussian mixture models) can then be applied in the transformed space where independence assumptions are more valid. In a similar spirit, we use LDA to reduce the vocabulary size of a bag of words representation – with typically tens of thousands of words – to a small number of topics. Models developed for data expressed in a topic representation can thus more justifiably make independence assumptions. We observe via experiment that LDA topics are superior to smoothed word counts as features to our model.

We comment that while our preliminary experiments led us to select these particular feature representations, the probabilistic model in the next section is applicable to any feature type that has been transformed into a vocabulary or topic representation.

3.2 The Model

Our model consists of a principled combination of generative and discriminative elements. We model the features of N_i images with corresponding web text. Let h be a binary random variable indicating if a given image retrieved from a query should indeed be be associated with a user defined concept of interest, such that $h \in \{$"relevant","not relevant"$\}$. Let a represent a binary random variable for a match of the query string with a sub-string in the image filename. Let b represent another binary random variable, indicating if there is an exact substring match in the html url.

We treat the raw words in a region surrounding the anchor text for the image as a set of random variables which we analyze using an LDA topic model. From the topic representation we are thus naturally able to characterize each document as consisting of N_w draws from a smaller vocabulary of *topic words* $W = \{w_1, w_2, \ldots, w_{N_w}\}$. SIFT features are quantized into a visual vocabulary as discussed in Section 3.1. Let the collection of quantized SIFT features for a given image be represented as a set of visual words $V = \{v_1, v_2, \ldots, v_{N_v}\}$.

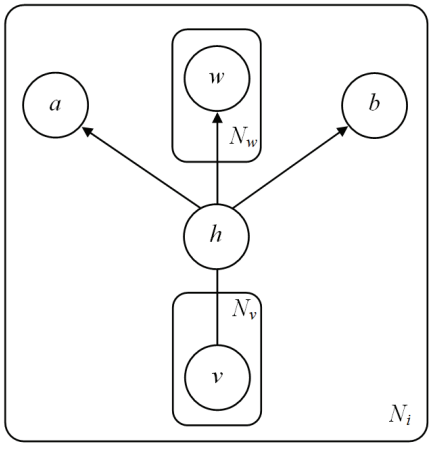

Figure 1: A graphical model with both generative (directed arrows) and discriminative elements (undirected links).

We use X to denote the set of all random variables other than h for which our model encodes a joint probability distribution, $X = \{a, b, W\}$. As we shall soon see, our model encodes a generative distribution on X when conditioned on V. With the above definitions in place we use the model illustrated in figure 1 given by

$$\begin{aligned} P(X, h|V) &= P(a, b, W|h)P(h|V) \\ &= P(a|h)P(b|h)P(W|h)P(h|V). \end{aligned} \tag{1}$$

For $P(h|V)$ we use a simple conditional random field

$$P(h|V) = \frac{1}{Z(V)} \exp\left\{ \sum_{k=1}^{K} \lambda_k f_k(h, V) \right\}. \tag{2}$$

For (2) we define k feature functions $f_k(h, V)$. In our experiments, function k evaluates to the number of counts for visual word k for each of our N_v possible visual words. For $P(a|h)$ and $P(b|h)$ we use binary distributions, and for our topical text words $P(W|h)$ we use a discrete distribution with the factorization

$$P(W|h) = \prod_{i=1}^{N_w} P(w_i|h). \tag{3}$$

Let the complete parameter space be denoted as $\mathbf{\Theta} = \{\lambda, \theta_a, \theta_b, \theta_w\}$, corresponding to a partitioning of parameters for each of the respective components of the underlying graphical model.

3.3 Estimation with Labeled Data

When we have hand specified labels for h, our approach decomposes into separate estimation problems for the generative and discriminative components of the model. Our objective for labeled data is

$$\log P(a, b, W, h|V) = \log P(a, b, W|h) + \log P(h|V). \tag{4}$$

The generative component of our model involves

$$\begin{aligned} \mathcal{L}_{X|h} = \log P(a, b, W|h) &= \sum_{i=1}^{N_w} \log P(w_i|h) \\ &+ \log P(a|h) + \log P(b|h), \end{aligned} \tag{5}$$

affording closed form updates for the parameters of conditional distributions for $P(X|h)$ easily computed from the standard sufficient statistics for Bernoulli, Discrete and Multinomial distributions. In contrast, parameter estimates for discriminative or undirected components of our model are obtained iteratively by gradient descent using

$$\begin{aligned} \mathcal{L}_{h|V} = \log P(h|V) &= \sum_{k=1}^{K} \lambda_k f_k(h, V) - \log Z(V) \\ &= \lambda^T \mathbf{f}(h, V) - \log Z(V) \end{aligned} \tag{6}$$

$$\frac{\partial \mathcal{L}_{h|V}}{\partial \lambda} = \mathbf{f}(h, V) - \sum_{h} \mathbf{f}(h, V) P(h|V). \tag{7}$$

3.4 Estimation with Unlabeled Data

We present a hybrid expectation maximization / expected gradient procedure to perform learning with unlabeled data. For unlabeled data we wish to perform optimization based on the marginal probability

$$p(a, b, W|V) = \sum_h P(a, b, W|h)P(h|V). \quad (8)$$

Our objective is thus

$$\mathcal{L}_{X|V} = \log \sum_h P(X|h)P(h|V), \quad (9)$$

and parameter estimation involves

$$\frac{\partial \mathcal{L}_{X|V}}{\partial \theta} = \sum_h P(h|X, V) \frac{\partial}{\partial \theta} \big[\log P(X|h) + \log P(h|V) \big]. \quad (10)$$

We thus perform our optimization by computing an expectation or E-step followed by: (1) a single closed form maximization or M-step for parameter estimates involving variables in X

$$\Theta_{x_j|h} = \frac{\sum_{i=1}^{|D|} P(h_i|V_i, X_i)N(x_{j,i})}{\sum_{i=1}^{|D|} \sum_{s=1}^{|X|} P(h_i|V_i, X_i)N(x_{s,i})} \quad (11)$$

and (2) an iterative expected gradient descent optimization (until local convergence) for the discriminative component of the model

$$\frac{\partial \mathcal{L}_{h|V}}{\partial \lambda} = \sum_h \mathbf{f}(h, V)P(h|X, V) - \sum_h \sum_X \mathbf{f}(h, V)P(h, X|V). \quad (12)$$

$N(x_{j,i})$ represents the number of occurrences of word x_j in document i. The complete optimization procedure consists of iterations of: an E-step followed by an M-step and an expected gradient optimization, repeating these steps until the likelihood change over iterations is within a relative log likelihood change tolerance of .001.

When we have a mixture of very small number of positive example images among many unlabeled instances, semi-supervised techniques can be sensitive to the precise selection of labeled elements [Druck *et al.*, 2007]. In our approach here whenever we have a mixture of labeled and unlabeled data we train our model on the labeled data first to obtain initial parameter estimates prior to a more complete optimization with both the labeled and unlabeled data. In our experiments we also use a random sampling procedure to simulate user specified labels.

4 Experiments

In the following section we describe experiments with our model on a web dataset themed by musical instruments. We cover the acquisition and processing procedure for the dataset, and evaluate model performance with respect to different visual queries and varying levels of supervision. We show that each major component of our model contributes to improved performance over baseline methods. Finally, we show that our method can be used to enhance object recognition performance on the challenging Caltech-256 dataset.

	Instances	True Instances
french horn	549	112
harp	531	93
harpsichord	394	113
piano	503	59
saxophone	521	102
timpani	480	37
tuba	487	67
violin	527	130
xylophone	495	58

Table 1: Musical instruments web dataset.

4.1 Web Dataset Acquisition

We wish to evaluate our model on a dataset consisting of unprocessed images and html text returned from web image search. We are unaware of a standard dataset with these characteristics, so we construct a new manually labeled dataset from Google Image Search results. Table 1 lists our choice of query words which follow a musical instruments theme. Careful consideration is needed to properly annotate the set; we define the following list of rules for consistent manual labeling with respect to $h \in \{$"relevant","not relevant"$\}$:

- Image is a clear, realistic depiction of the most common form of the object defined by the query word.
- Query word is the focus of the image, i.e. the object of interest must be at least as prominent in the image as any other object or theme.
- Reject images containing more than 3 instances of the object of interest. Repeated instances of an object begin to resemble a texture.
- Reject cartoon images and abstract art.
- Reject if the context surrounding the object seems bizarre or impossible.
- Include images of objects which are sufficiently similar as to fool a human observer (e.g., we include images of mellophones with the french horn class).

4.2 Filtering for Visual Concepts

We run a series of repeated experiments for each of the classes listed in table 1. For each repeat under a given class we vary the training set (corresponding to example images selected during user intervention) by randomly choosing 5 images from the pool of true-class images. We then run the hybrid semi-supervised algorithm until convergence, which usually occurs within 30 iterations.

Figure 3 shows the average precision-recall curve for each class given by the model and by the Google rank. We note that our model outperforms the original search engine ranking for every class, sometimes significantly so. Search engine results have the desirable property of high precision among the top-ranked elements, and we see the same effect in our model rerankings.

Figure 2: Left: top-ranked images by our algorithm. Right: top images returned by Google for "binoculars" query.

	G	text	vis	no-EM	EM
french horn	0.70	0.56	0.82	0.80	**0.92**
harp	0.60	0.58	0.40	**0.70**	0.62
harpsichord	0.60	0.78	0.90	0.90	**0.94**
piano	0.70	0.68	0.50	**0.80**	**0.80**
saxophone	0.70	0.80	0.78	0.78	**0.90**
timpani	0.60	0.56	**0.90**	0.70	0.86
tuba	0.50	0.56	0.90	0.94	**0.98**
violin	0.70	0.34	**0.86**	0.72	**0.86**
xylophone	0.40	0.74	0.68	**0.80**	0.78
μ	0.61	0.62	0.74	0.79	0.85
σ	0.10	0.14	0.18	0.09	0.09

Table 2: Top-10 accuracy for each model variant and class. Top-10 accuracy represents the proportion of images correct among the top-10 ranked images, averaged over multiple trials with different training sets. G: Google Rank. text: text-only model. vis: visual-features-only model. no-EM: full model with a single EM step. EM: complete model.

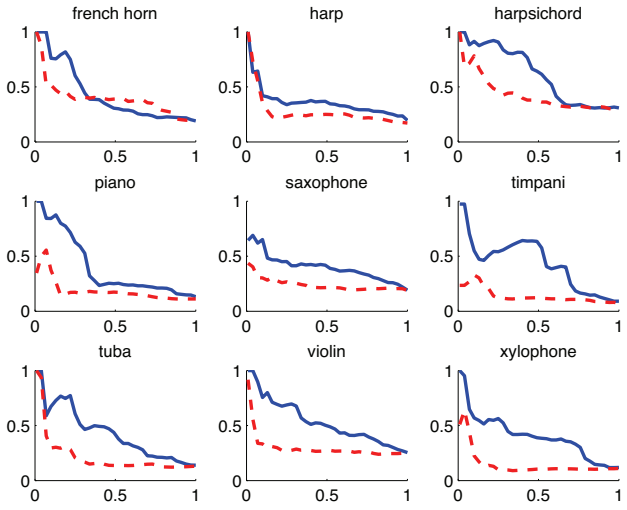

Figure 3: Precision-Recall curves for each of the categories in our web dataset. Solid blue curves:our model; dashed red curves: Google rankings.

4.3 Comparisons to baseline methods

It was shown in [Berg and Forsyth, 2006] that for some image search queries, text features are more reliable predictors for classification, while for other classes visual features are more useful. This makes intuitive sense and we wish to validate this argument for our model. We compare the performances of a number of variants of our model to judge the merits of individual components.

The variants we tested include the model without the hybrid expectation maximization procedure, a portion of the model using only text features, and a portion of the model using only image features. In these tests we again use 5 randomly chosen positive examples for the labeled training sets.

Our findings agree with those of [Berg and Forsyth, 2006] where for some classes text features are more important while for others visual features are more useful. The combination of features in our model yields better performance in almost every case. Additionally, our hybrid expectation-maximization procedure further improves results. Table 2 lists the accuracy among the top-10 ranked images for each method and each class.

4.4 Training Visual Classifiers with Web Data

Here we explore the benefits of augmenting traditional image classification experiments with additional data learned from the proposed model. We fetch hundreds images and associated text from Google for each category name in the Caltech-256 dataset. Using a small number of Caltech-256 images for initialization, we train a separate model for the web data of each category. The top 30 ranked web images per category are added to the pool of initialization training images, and 1-vs-all maximum entropy classifiers are trained on this data. A sampling of the top images learned by our model is presented in figure 2, where our results are compared to Google ranked results. Figures 4 and 5 compare performance of these augmented classifiers to classifiers trained on the initialization training data alone. We see that the addition of images learned from our model enhances classifier performance, particularly in the case of small training set sizes. Although our simple bag-of-words representation and maximum-entropy classifiers do not match state of the art performance on Caltech-256, the inclusion of clean web images learned by our model should benefit a wide range of classification techniques.

5 Conclusions

We have presented a method for constructing visual classifiers from the web using minimal supervision. Our approach is built upon a novel probabilistic graphical model which

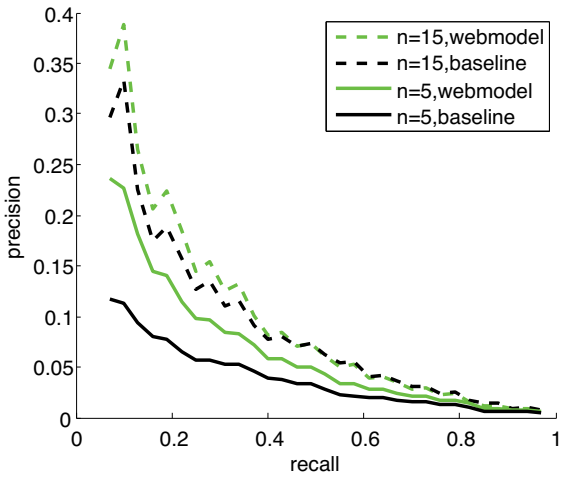

Figure 4: Average Precision-Recall on Caltech 256. n is the initialization positive training set size.

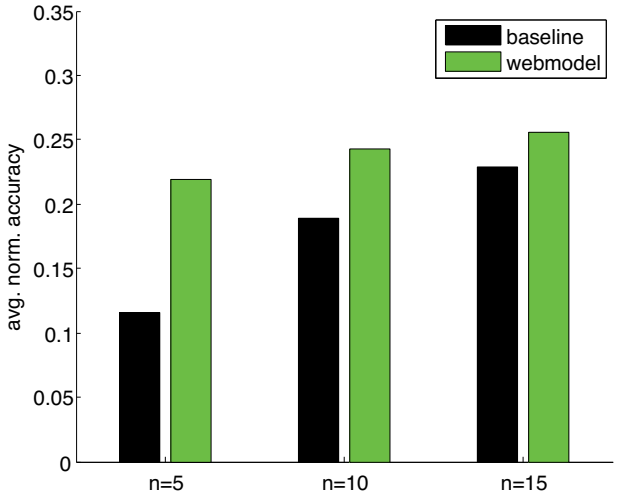

Figure 5: Average normalized accuracy on Caltech-256.

combines image features and text features from associated html documents. We introduced a hybrid expectation maximization / expected gradient procedure and showed that this semi-supervised approach gives better performance than a number of baseline tests. The model was applied to a dataset of musical instruments collected from the web, and the resulting rerankings significantly improved upon rankings given by Google image search. Top images from the reranked data have correct labelings and improve performance when training classifiers for object recognition tasks.

References

[Berg and Forsyth, 2006] TL Berg and DA Forsyth. Animals on the Web. *CVPR*, 2, 2006.

[Blei *et al.*, 2003] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *JMLR*, 3(5):993–1022, 2003.

[Dempster *et al.*, 1977] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[Druck *et al.*, 2007] Greg Druck, Chris Pal, Jerry Zhu, and Andrew McCallum. Semi-supervised classification with hybrid generative/discriminative methods. In *KDD*, 2007.

[Fergus *et al.*, 2005] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. *ICCV*, 2, 2005.

[Lasserre *et al.*, 2006] J. Lasserre, C. M. Bishop, and T. Minka. Principled hybrids of generative and discriminative models. In *CVPR*, 2006.

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

[Lowe, 1999] D.G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 2:1150–1157, 1999.

[Lu *et al.*, 2000] Y. Lu, C. Hu, X. Zhu, H.J. Zhang, and Q. Yang. A unified framework for semantics and feature based relevance feedback in image retrieval systems. *Eighth ACM international conference on Multimedia*, pages 31–37, 2000.

[McCallum *et al.*, 2006] A. McCallum, C. Pal, G. Druck, and X. Wang. Multi-conditional learning: Generative/discriminative training for clustering and classification. *Proceedings of 21st National Conference on Artificial Intelligence (AAAI)*, 2006.

[McCallum, 2002] A. McCallum. Mallet: A machine learning for language toolkit. http://mallet.cs.umass.edu, 2002.

[Ng and Jordan, 2001] A. Ng and M. Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *NIPS*, pages 841–848, 2001.

[Nowak *et al.*, 2006] E. Nowak, F. Jurie, and B. Triggs. Sampling strategies for bag-of-features image classification. *Proc. ECCV*, 4:490–503, 2006.

[Ponce *et al.*, 2006] J. Ponce, TL Berg, M. Everingham, DA Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, BC Russell, A. Torralba, et al. Dataset Issues in Object Recognition. *Toward Category-Level Object Recognition. LNCS*, 4170, 2006.

[Rui *et al.*, 1998] Y. Rui, TS Huang, M. Ortega, and S. Mehrotra. Relevance feedback: a power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE*, 8(5):644–655, 1998.

[Salakhutdinov *et al.*, 2003] Ruslan Salakhutdinov, Sam T. Roweis, and Zoubin Ghahramani. Optimization with em and expectation-conjugate-gradient. In *ICML*, 2003.

[Schroff *et al.*, 2007] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[Torralba *et al.*, 2007] A. Torralba, R. Fergus, and W. T. Freeman. Tiny images. Technical Report MIT-CSAIL-TR-2007-024, CSAIL, MIT, 2007.