Reducing correspondence ambiguity in loosely labeled training data

Kobus Barnard
University of Arizona
Tucson Arizona
kobus@cs.arizona.edu

Quanfu Fan
University of Arizona
Tucson Arizona
quanfu@cs.arizona.edu

Abstract

We develop an approach to reduce correspondence ambiguity in training data where data items are associated with sets of plausible labels. Our domain is images annotated with keywords where it is not known which part of the image a keyword refers to. In contrast to earlier approaches that build predictive models or classifiers despite the ambiguity, we argue that that it is better to first address the correspondence ambiguity, and then build more complex models from the improved training data. This addresses difficulties of fitting complex models in the face of ambiguity while exploiting all the constraints available from the training data. We contribute a simple and flexible formulation of the problem, and show results validated by a recently developed comprehensive evaluation data set and corresponding evaluation methodology.

1. Introduction

There has been much recent interest in learning to recognize semantic elements from training data which has multiple labels (e.g., images with associated text). For example, one version of the Corel™ data set has nearly 40,000 images with 4-6 keywords. A second example is news photos with captions, available in large quantities on the web [6]. By its nature, such data has correspondence ambiguity, because which label goes with which image element (if any), is not known. Nonetheless, a number of recently proposed methods have been developed which can learn to recognize and label regions or, more simply, annotate images with appropriate words under these conditions [1, 3, 5, 10, 11, 18, 20, 21, 23, 29].

By necessity, all these methods are trained on the annotation task. Region labeling can be learned collaterally provided that the model or classifier produces better annotations as region labeling improves. However, notice that achieving reasonable performance on image annotation does not require breaking correspondence ambiguity. For example, if horses and grass co-occur frequently, then learning to guess either appropriately is a reasonable strategy. However, we suggest that such fortuitous results only go so far, even in the context of retrieval, and hence we need to ensure that methods and corresponding evaluation strategies include localization. As well argued with others [2, 14, 15, 22], for retrieval to be really useful, we need much better understanding of image semantics, which implies spatial localization.

Most region labeling approaches either fit a simply statistical model to the training data [3, 5, 10, 11], or build classifiers for each label despite the ambiguity in the training set (i.e. "multiple instance learning" [1, 23, 29]). The first approach has the difficulty that the form of a good generic model is not known, and even simple models are difficult to learn effectively.

In the second approach, classifiers for each word are learned independently, which both ignores some available information and implies a substantial computational burden with large vocabularies. The fact that a region should have one (or perhaps a few) labels must be dealt with separately which has not been adequately addressed so far. In particular, we are not aware of a multiple instance learning approach which exploits the notion that, to the extent that a region is believed to be associated with a certain label (e.g. "tiger"), that region should not be labeled as something else. This consideration is one manifestation of what we will refer to as exclusion reasoning, and is embodied in probabilistic methods which attempt to distribute label probabilities subject to the constraint that they sum to one.

A second manifestation of exclusion reasoning is that assigning a label to one region should reduce the expectation that other regions should receive that label. One formulation of this is to assume that every label has at least one region associated with it. We are not aware of any approaches that explicitly to exploit this notion, but it has the potential for being very helpful in learning rare words. Figure 1 illustrates some of the above ideas.

Other sources of information include spatial context (a brown blob in the sky is more likely to be a bird than on inside a building); adjacency cues (two different adjacent regions are often parts of the same semantic entity); and common configuration (car wheels and other car parts located near each other reinforce the notion of car. All these cues are useful, but each one adds complexity to models, making fitting even more difficult.

1-4244-1180-7/07/$25.00 ©2007 IEEE

1.1. Our approach

To address these problems, we propose a substantively different approach. We suggest developing strategies to first label the training data as best as possible, and then deal with the issues of classification and/or building models suitable for inference on new images. In other words, we want to push the loosely labeled data towards truly labeled data.

The advantages of a large scale, reasonably labeled training set should be clear. Given better labels, we can develop a variety of processes that learn to classify image semantic entities based on a variety of features, spatial context, and complex spatial and part models. Importantly, these processes can be quite distinct from those used to reduce correspondence ambiguity.

We propose bootstrapping the learning process in this way because improving the training data labeling is an easier problem than learning models for all semantic entities. Thus it should not be solved as a consequence of solving the more complex problem. In particular, there are a number of ways to exploit the hidden supervisory information available in loosely labeled training data that are awkward to integrate completely into the fitting process of generic models.

One difficulty in labeling regions, especially with a modest feature set, is that the process is inherently ambiguous. A smooth red region that is part of a red car can be very similar to regions that are parts of toys or flowers in other images. Assuming that a reasonable feature set cannot distinguish between them, higher order cues such as context are needed—but we need large training data sets to learn these. Interestingly, in the case of loosely labeled data, we have a good chance of labeling the regions without having a contextual model because we can limit the choices to the set of labels. In short, the labels can identify likely consistent contexts for us.

This is important because learning a context model simultaneously with the labels can be substantively more complex and difficult. Examples of existing approaches that attempt to incorporate context include various forms of document level clustering which essentially provide priors over the region feature models [3, 8, 28], and augmenting translation models with a more explicit model for contextual relationships among the labels [9]. The complexity of these models underlies our main tenet that it may be better to address the ambiguity first—in short, separate the correspondence ambiguity problem and the inference problem.

The main strategic elements of our approach are:

- We use a weak model for linking features to semantics because one of the main difficulties is that we do not know what models are good for which image entities. The specific method proposed here is built upon a simple affinity matrix that is tuned to approximate the well defined semantic notion that the affinity between two regions should approximate the probability that they have the same label.
- We only exchange information between regions with similar features *provided* that they have compatible words (*implied context*). Similar regions can have different labels, and thus they need to be distinguished by higher level analysis. Here we are focused on developing the training data to simplify learning the visual context and second order structure to make that possible.

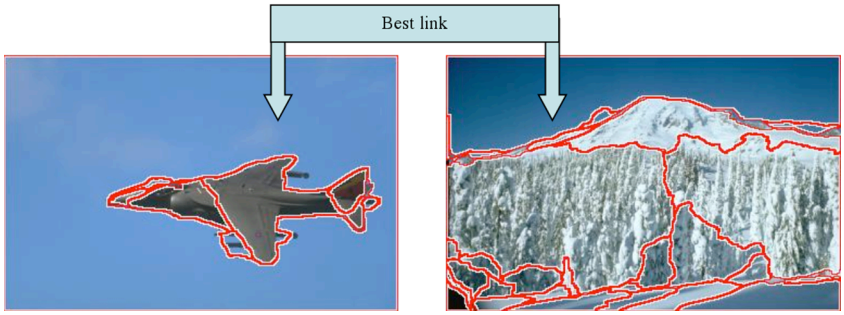

Figure 1. A friendly example illustrating some of the ideas in this paper. The associated words are shown below each image. Regions that are close together in feature space will propagate shared words, of which there is only one ("sky") in this example. Assuming such evidence for sky in the left image, exclusion reasoning means that we have some chance to get the other labels correct, as the remaining words will be assigned to the remaining regions that are most unlike sky. In this work we assume (incorrectly) that the synonyms "plane" and "jet" and the specific term "harrier" all refer to different regions. This problem can be mitigated with language processing (e.g. [19]), which would allow us to establish the labels for all regions in the left hand image.

- We employ exclusion reasoning which is essentially a matching constraint. In particular, we assume that every image has at least one region for every label. This means that we can hypothesize regions for relatively rare words. Candidate regions for the excluded label are ones that have less support to be associated with other words.

1.2. Related work

In addition to the body of work on image annotation mentioned above, this work is related to propagating labels on manifolds [30] and spectral clustering (e.g. [26, 27]). There are several interesting differences that arise in our domain. Trivially, we don't learn from either truly labeled data, nor truly unlabeled data. Instead, we are focused on reducing the ambiguity in the labels, analyzed as a group. Also, since we consider generalizing to new data a second step, over fitting issues are less problematic. Another difference is that we do not simply assume that similar regions should be labeled the same, and ones that are dissimilar should be labeled differently, as there are too many deviations from this. Finally, perhaps the largest difference between our algorithm and other related ones is that the main driver for the solution are the constraints.

2. The algorithm

We first construct an affinity matrix, A, which is assumed to estimate the probabilities that all pairs of regions under consideration have the same label. This entails a simple expression on a the region label probability matrix, U. We estimate a solution to this equation, subject to the exclusion reasoning constraints. The details follow.

2.1. The affinity matrix

It is common to embody a notion of feature similarity among N items by the entries of an N by N symmetry affinity matrix. Interestingly, it is less common to attempt to specify what semantics the entries should specify for a particular algorithm. For our problem, we propose that the most natural definition is:

A_{ij} = Probability that region i and j have the same label

For convenience, we assume that A can be estimated by the form:

$$A_{ij} \propto \hat{A} = \begin{cases} \exp(-\sum_k \left(x_k^i - x_k^j\right) w_k \left(x_k^i - x_k^j\right) & i \neq j,\ i \text{ and } j \text{ share words} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

The subscript, k, indexes over D features which have weights, w_k, and we do not use the diagonal elements of A. To set the parameters, which is a form of feature selection tuned to the task, we use ground truth data held out from testing. In particular, we compute a value encoding the extent that two regions should be labeled the same, taking into account that both typically overlap multiple semantic regions due to segmentation problems. This information can be computed from the score matrices delivered by our ground truth methodology [4]. For each such pair, we have:

$$-\log\left(p_{ij}\right) = \sum_k \left(x_k^i - x_k^j\right) w_k \left(x_k^i - x_k^j\right) \quad (2)$$

This gives a linear equation in the D variables w_k. We solve this in the least squares sense using many region pairs. We weight the equations so that the errors in p_{ij} space are uniformly considered, as opposed to simply solving the system of equations with uniform equation weights which minimizes error in log space.

While (1) this is a relatively crude way to compute A, by being specific about its semantics, we are in good position to seek better ways to estimate it.

2.2. The main equation

We consider our unknown, U, to be a matrix of dimension N by W where N is the number of regions in the training data, and W is the vocabulary size. A row, **u**, of U, is the probability distribution over labels for that region.

The probability that two different regions, i and j, have the same label, assuming independence where needed, is then simply the marginalization over words, easily represented by the dot product:

$$p_{ij} = \mathbf{u}_i \bullet \mathbf{u}_j \quad (3)$$

Taking all such pairs together, in conjunction with our assumed semantics for the affinity matrix we get:

$$UU^T = A = \alpha \hat{A} \quad (4)$$

where α is a constant that is estimated given $\hat{A}$ and an estimate for U. Our goal becomes finding a good U for the above subject to constraints.

2.3. The constraints

The probabilistic interpretation immediately suggests the constraints:

$$0 \leq U_{iw} \leq 1 \quad \text{and} \quad \sum_w U_{iw} = 1 \quad (5)$$

This second constraint embodies the notion of weak exclusion reasoning that assumes that each region should have one label. Clearly this assumption can be violated in practice. Our approach leaves room for alternative constraints, but here we only experiment with this form of weak exclusion reasoning.

We constrain U so that words that do not occur in the observed labels never have positive probability. The labels for the source image provide *implied context* for the region with specific features. For example, a flat red region is

reasonably likely to be a car if the image has the keyword "car", and is reasonably likely to be part of a flower if the image has keyword "petal" or "flower". This is embodied by the simple constraint:

$$U_{iw} = 0 \quad \text{if image for region } i \text{ does not have } w \qquad (6)$$

This does not allow for words that are missing from the annotation. A planned extension is to estimate the probability that a word is missing, and use that probability to help set the bound above.

The final constraint in our current implementation is that each word should be associated with at least one region unless there are more words than regions for that image. This implements the second manifestation of exclusion reasoning discussed above. This is encoded by the constraint:

$$\sum_{i \in I} U_{iw} \geq \min\left(1, \frac{num\_regions\_in\_I}{num\_words\_for\_I}\right) \qquad (7)$$

2.4. Estimating the label probabilities

We attempt to solve the main equation (4) in the least squares sense. Unfortunately, the objective function is non-convex. Further, for a reasonable size problem, the number of regions, N, is quite large. Thus we do not know of a practical method to find the global optimal for it. Fortunately, the problem has substantive structure that allows relatively efficient constrained gradient descent [7]. In particular:

- Regions only interact if they are from images with shared words.
- Because only entries of A that are from pairs of regions whose images share words, and many of those non-zero entries are small, and we can use a sparse representation for A.
- All the constraints are localized on an image bases.

We compute a step in the direction of the negative gradient of:

$$f(U) = \sum_i \sum_j \left(A_{ij} - \mathbf{u}_i \bullet \mathbf{u}_j\right)^2 \qquad (8)$$

so that the constraints are not violated.

In more detail, for a given region, the gradient of (8) is given by:

$$\frac{\partial f(U)}{\partial \mathbf{u}_i} = 4 \sum_j \left(A_{ij} - \mathbf{u}_i \bullet \mathbf{u}_j\right) \mathbf{u}_j \qquad (9)$$

We project these vectors onto subspaces implied by the equality constraints in (5) for each image. We further constrain the step size for each component so that the inequality constraints hold. Once we take a step, we verify that the proposed step will in fact decrease the value of $f(U)$, and back off it this is not the case. Finally, to begin the process, we initialize U so that all words from the image for each blob have equal non-zero probability, and the ones that do not occur have zero probability.

Finally, we update α in (4) as we solve for U, but it becomes stable after a small number of iterations.

Regardless of how this optimization problem is approached, it is important to realize that it is relatively sparse, and the sparseness likely needs to be exploited for real data set sizes. In particular, while there are nominally of the order of N by W variables, where N is the number of regions and W is the vocabulary size, all U_{ij} values for words that do not occur in annotation are implicitly zero, and do not need to be considered at all.

2.5. Classification of new data

Our general strategy is to use the better labeled training data as input to any of a variety of learning strategies. Since the above algorithm is focused on labeling the training data, it is not necessarily the ideal method for labeling new data. Nonetheless, we can use the machinery to do so, and we report results below.

The baseline approach for labeling images without labels is as follow. We initialize the probability distribution over the labels to the empirical distribution over the test data. We then run the gradient descent to improve the labels, but with only the new (test) images factoring into the computations. Further, and the constraints (6) and (7) are not helpful and are omitted. This prediction method is essentially non-parametric, with the process requiring (and exploiting) all training data.

2.6. Learning and using context

Given better labeled data, we can compute statistics for the spatial arrangements of regions that can be used to improve the labeling of test data. For example, if a brown region is adjacent to many "sky" regions, the probability that the region is "bird" should be increased. A sophisticated spatial semantic model is beyond the scope of this paper. However, we experimented with the following simple one.

Using the improved labeled data, we compute the empirical probability distribution for the semantic information available to due adjacent regions by:

$$p_{adj}(w_i, w_j) \propto \sum_m \sum_{\text{n} \in \text{adjacent(m)}} U_{mw_i} U_{nw_j} \qquad (10)$$

To use this information to improve test data labeling, we assume that the spatial information is independent of the feature based labeling. Then, given a feature based labeling for a region, the above is used to determine the word probability distribution gained from the spatial information by:

$$p_i(w \mid \text{adjacency}) = \sum_{j \in adjecent(i)} \sum_v U_{jv} p_{adj}(w \mid v) \qquad (11)$$

Because we assume independence, we simply multiply the feature based distribution with the above spatial based one, and renormalize.

3. Experiments

We consider two tests of our approach. First, and foremost, we examine the task of labeling the training data, as this is the focus of this work. Second, we evaluate the performance on held out data, with and without the simple context model just described.

To evaluate region labeling performance we exploit a recently developed data set and evaluation methodology [4]. In that work, we labeled human segmentations for 1014 Corel™ images from a different study [24] with words from the WordNet ontology [17, 25]. Further, we developed a methodology to map labelings corresponding to arbitrary segmentations and relative to any vocabulary. For example, the methodology provides a principled method for scoring the use of a more general word (e.g. "cat") versus a more specific one (e.g. "tiger). Thus we are able to automatically evaluate region labeling performance on those images.

That work further defines two measures of localized semantic performance: "range of semantics identified" and "frequency correct". A simple example will clarify the difference. Consider two algorithms, one which reliably identifies tigers, but nothing else, and a second one which reliably identifies sky, but nothing else. By the first notion, these two algorithms have the same performance (one semantic entity). By the second notion, the second algorithm performs better because sky is much more common, and thus a count of correctly identified entities over a reasonable test set will be higher.

Despite the fact that the Corel™ data has been noted as been somewhat easy for studying some related tasks [12], we argue that the particular task under consideration here is very difficult, even on the Corel™ data. First, perfect results require perfect semantic segmentations. Since our segmentations are far from perfect, we provide values for the maximum possible scores given the segmentations and the training set vocabulary. Second, optimal performance requires identifying localized semantics at the most specific level given the vocabulary (e.g., using "tiger" instead of "cat"). Hence we do not expect high levels of absolute performance.

3.1. Baseline

Any region labeling method can be used to label the training data, and typically the performance of such a method will be greatly improved by restricting word prediction to the associated words. We use this approach applied to a multi-modal mixture model (MMMM) [3] as a baseline for performance evaluation. This model assumes that that images and associated text are generated by choosing one or more concepts (latent factors), l, from a prior distribution, $P(l)$, and then by generating regions and associated words conditionally independent given the latent factors. Thus, the joint probability of an image region and a word is expressed by

$$P(w,r) = \sum_l P(w \mid l) P(r \mid l) P(l) \qquad (12)$$

where w denotes a word, r denotes a region, l indexes latent factors, $P(w \mid l)$ is a probability table over the words, and for the blob model, $P(r \mid l)$, we use a Gaussian distribution over features with diagonal covariance matrices. For the experiments we set the number of factors to 500.

The above equation provides a word posterior distribution for a region, but to train the model we need to specify how to generate the observations (i.e., image annotations). Here we assume each observed region provides a posterior over the latent factors. These distributions are simply summed up, to provide overall mixing weights for the components $P(w \mid l)$. The resultant posterior distribution over words is then sampled to generate the words for the image. The model is trained with the expectation maximization (EM) algorithm [13], using missing values to represent the hidden factor responsible for each word and region. This is the proposed DEPENDENT version of the MMMM [3]. We adopt that name here.

We also experimented with the CORRESPONDENCE version [3] of the MMMM because it strongly embodies the notion of exclusion reasoning during training, because it attempts to fit a 1-1 correspondence between image regions and words. Thus it seems that a comparison with our approach has merit. This form of the MMMM only differs from the DEPENDENT version during training. Inference is still done by (2).

3.1.1 Restricted word prediction

Once the model is trained, then the probability distribution over words for a given region, $P(w \mid r)$, is easily computed using (12). However, to label training data, we take one further step. We only allow the model to predict words that are observed. Operationally we set the probability of the other words to zero, and then renormalize .

The multi-modal mixture model approach has enjoyed good success at predicting common visual words for both for images and for regions. In fact, the most careful study of region labeling that is available [4] suggests that it is roughly the state of the art on this task. When combined with restricted word prediction, it is a stiff baseline method for training set labeling.

3.2. Experimental methodology

The 1014 manually labeled images came from 275 Corel™ CD's. We constructed a data set from the images from those CD's. We segmented these images using normalized cuts [27], and extracted features similar to those used by Barnard et al. [3]. A few images were

omitted due to various problems, leaving a pool of 27,128 images. We then produced sixteen subsets of the data. Here we divided the CD's into sixteen roughly equal parts to ensure that there were some training images from the same CD for each the manually labeled images. For the experiments on labeling the training data, the manually labeled data were included as part of the training data; otherwise they were held out. We restricted the vocabulary to words that occurred with at least 20 images. The vocabulary size ranged from 40 to 80 words over the sixteen sets.

3.3. Results

The results in Table 1 show that approach developed here can do well labeling the data in comparison to the baseline. We do particularly well in semantic range performance, which is satisfying since the algorithm was designed to do well on less common words which are harder to learn. We also exceed the restricted MMMM in frequency correct performance by a modest amount. In training, the MMMM is rewarded for frequency correct (on an image basis), and thus tends to learn to predict common words. In fact, in the Corel^{TM} data, substantively

Algorithm	Freq	Range
Empirical distribution	2.05 (1.8)	0.74 (0.3)
Dependent MMMM	4.89 (1.2)	2.50 (0.5)
Correspondence MMMM	3.80 (1.4)	2.25 (0.6)
Restricted dependent MMMM	8.00 (1.3)	5.65 (1.1)
Restricted correspondence MMMM	7.44 (1.5)	5.37 (1.1)
Result using the initial estimate	5.88 (0.7)	5.26 (1.3)
The main method proposed here	**8.67** (0.9)	**7.40** (1.0)
Theoretical maximum	22.3 (3.3)	20.3 (3.2)

Table 1. Performance on training data. This table shows quantitative region labeling results for the method proposed in the text compared with two versions of the multimodal mixture model (MMMM). These are averages over 16 splits of 27,128 images including 1014 semantically region labeled images among them. The numbers in parentheses is an error estimate based on the variance of results over the splits. As discussed in the text this is the performance on training data. The maximum achievable result (last row) with the ground truth scoring system is a function of the segmentation quality. These results show a substantive improvement over the stiff base line (the restricted dependent version of the multi-modal mixture model) in the case of the semantic range measure, and a modest improvement in the case of frequency correct. Increasing the frequency correct number in an absolute sense by improving the labeling of less common words is difficult on this data set because it is heavily weighted towards common words.

exceeding the performance of even the empirical distribution on an absolute scale is difficult because a few words (e.g. "sky", "water", and "people") dominate. Because of this, improving performance on rare words only has a small impact on performance number. Thus the observed increment, as corroborated by the excellent semantic range performance, is very promising.

The results provided in Table 2 suggest that our approach can also provide good region labeling results on new data. In fact, when combined with the relatively naïve context model, the performance significantly exceeds that of the baseline, which is close to the state of the art as reported elsewhere [4]. Notice that the context model helps semantic range performance more than frequency correct, where we observed only a modest gain. This makes sense. When there are lots of examples, some are likely to be close to the region under consideration. When there are fewer examples, features become more ambiguous, and context becomes more relevant.

Finally we comment that the CORRESPONDENCE version of the MMMM did not perform as well as we expected on the semantic range task. We hypothesize that perhaps the form of exclusion reasoning implemented in that algorithm is perhaps too strong. Further work is required to determine the details.

Figure 2 shows some region labeling results. While such qualitative results cannot replace the large scale evaluation, it is easy to find examples where our method does better, consistent with the fact that we label a substantively larger range of regions better than the baseline.

Algorithm	Freq	Range
Empirical distribution	2.17 (2.0)	0.74 (0.3)
Dependent MMMM	4.55 (1.2)	1.95 (0.4)
Correspondence MMMM	3.14 (1.7)	1.67 (0.6)
The main method proposed here	5.25 (1.7)	2.69 (0.8)
The method with context model	**5.31** (1.5)	**3.50** (1.2)
Theoretical maximum	22.3 (3.3)	20.3 (3.2)

Table 2. Performance on data not used for training. This table provides classification results for regions from held out images from the ground truth. The numbers shown are averages over 16 splits, with 1014 semantically region labeled images held out in total. In fact, the test data is exactly the same as that for Table 1, except that here they were not also used for training. Thus the maximum possible score achievable is the same as in Table 1. The low numbers relative to this simply reflect the difficulty of the task. We observe that our approach also does a good job of labeling data not used in training, and that the context model helps, most noticeably in the case of the semantic range performance measure.

4. Discussion and future work

We have argued that there is merit in addressing the correspondence ambiguity in loosely labeled training data. In particular we advocate addressing this problem as a separate problem from learning models or building classifiers. The system that we have developed for doing this performed very well on the most important task that we set out for it, namely labeling a wider range of semantic entities correctly. This is important if we are to mine large, loosely labeled data for semantically meaningful visual patterns.

There are many ways that our approach could be improved, some of which have been mentioned in the text. To further improve labeling of the training data, we would like to investigate the benefit of modeling missing word probabilities estimated from the observed words. We also expect that reducing vocabulary redundancy using WordNet [16, 17, 25] will be helpful. Finally, we would like to investigate alternative embodiments of exclusion reasoning.

We are, of course keen to investigate the next step, namely building better recognizers from the better labeled data. Further, with our general approach, learning to use context and spatial configuration models is rendered much easier, and hence can be effectively explored.

References

[1] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multiple-Instance Learning," *Proc. Advances in Neural Information Processing Systems, 15*, Vancouver, BC, 2002.

[2] L. H. Armitage and P. G. B. Enser, "Analysis of user need in image archives," *Journal of Information Science*, vol. 23, pp. 287-299, 1997.

[3] K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching Words and Pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107-1135, 2003.

[4] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, and J. Kaufhold, "Evaluation of localized semantics: data, methodology, and experiments," University of Arizona, Computing Science, Technical Report, 2005, Available from http://kobus.ca/research/publications/IJCV-06/TR-05-08-revised.pdf.

[5] K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures," *Proc. International Conference on Computer Vision*, pp. II:408-415, 2001.

[6] T. L. Berg, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller, and D. A. Forsyth, "Names and Faces in the News," *Proc. Computer Vision and Pattern Recognition (CVPR)*, Washington D.C, pp. 848-854, 2004.

[7] D. P. Bertsekas, *Nonlinear Programming*, 2nd ed. Belmont. Mass.: Athena Scientific, 2003.

[8] D. M. Blei and M. I. Jordan, "Modeling annotated data," *Proc. 26th International Conference on Research and Development in Information Retrieval (SIGIR)*, 2003.

[9] P. Carbonetto and N. d. Freitas, "Why Jose can't read," *Proc. HLT-NAACL workshop on learning word meaning from non-linguistic data*, Edmonton, Alberta, pp. 54-61, 2003.

[10] P. Carbonetto, N. d. Freitas, and K. Barnard, "A Statistical Model for General Contextual Object Recognition," *Proc. European Conference on Computer Vision*, 2004.

[11] P. Carbonetto, N. d. Freitas, P. Gustafson, and N. Thompson, "Bayesian Feature Weighting for Unsupervised Learning, with Application to Object Recognition," UBC, Unpublished manuscript, 2002, Available from http://www.cs.ubc.ca/~nando/papers/shrinkage.pdf.

[12] O. Chapelle, P. Haffner, and V. Vapnik, "SVMs for Histogram Based Image Classification," *IEEE Transactions on Neural Networks*, vol. 9, 1999.

[13] 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. Series B (Methodological)*, vol. 39, pp. 1-38, 1977.

[14] P. G. B. Enser, "Query analysis in a visual information retrieval context," *Journal of Document and Text Management*, vol. 1, pp. 25-39, 1993.

[15] P. G. B. Enser, "Progress in documentation pictorial information retrieval," *Journal of Documentation*, vol. 51, pp. 126-170, 1995.

[16] C. Fellbaum, "WordNet: An Electronic Lexical Database," MIT Press, 1998.

[17] C. Fellbaum, P. G. A. Miller, R. Tengi, and P. Wakefield, WordNet - a Lexical Database for English, http://www.cogsci.princeton.edu/~wn.

[18] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic Image Annotation and Retrieval using Cross-Media Relevance Models," *Proc. SIGIR*, 2003.

[19] Y. Jin, L. Khan, L. Wang, and M. Awad, "Image Annotations By Combining Multiple Evidence & WordNet," *Proc. ACM multimedia*, 2006.

[20] V. Lavrenko, R. Manmatha, and J. Jeon, "A Model for Learning the Semantics of Pictures," *Proc. NIPS*, 2003.

[21] J. Li and J. Z. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 25, 2003.

[22] M. Markkula and E. Sormunen, "End-user searching challenges indexing practices in the digital newspaper photo archive," *Information retrieval*, vol. 1, pp. 259-285, 2000.

[23] O. Maron and A. L. Ratan, "Multiple-Instance Learning for Natural Scene Classification," *Proc. The Fifteenth International Conference on Machine Learning*, 1998.

[24] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," *Proc. International Conference on Computer Vision*, pp. II:416-421, 2001.

[25] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. J. Miller, "Introduction to WordNet: an on-line lexical database," *International Journal of Lexicography*, vol. 3, pp. 235 - 244, 1990.

[26] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," *Proc. Advances in Neural Information Processing Systems 14*, 2001.

[27] J. Shi and J. Malik., "Normalized Cuts and Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 888-905, 2000.

[28] E. B. Sudderth, A. Torralba, W. T. Freeman, and A. S. Willsky, "Learning Hierarchical Models of Scenes, Objects, and Parts," *Proc. ICCV*, 2005.

[29] Q. Zhang and S. A. Goldman, "EM-DD: An improved multiple-instance learning technique," *Proc. Neural Information Processing Systems*, 2001.

[30] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," *Proc. NIPS 16*, 2004.

Figure 2. Region labeling examples. Top left is the original image. Top middle is the result using the dependent model on the training data, and further restricted to output words that occur with the image. Top right is the algorithm developed in this paper, again labeling the training data. The bottom left image is the standard multi-modal mixture model. The bottom middle is the algorithm developed in this paper applied to data not used for training. The bottom right example is the same algorithm, with the added simple spatial context model. In general, we achieve better labeling than the multi-modal mixture model, but of course, there are still many errors, as the task is very difficult, even on an image with good segmentations such as this one.