Beyond Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers*

Abhinav Gupta and Larry S. Davis

Department of Computer Science
University of Maryland, College Park
{agupta,lsd}@cs.umd.edu

Abstract. Learning visual classifiers for object recognition from weakly labeled data requires determining correspondence between image regions and semantic object classes. Most approaches use co-occurrence of "nouns" and image features over large datasets to determine the correspondence, but many correspondence ambiguities remain. We further constrain the correspondence problem by exploiting additional language constructs to improve the learning process from weakly labeled data. We consider both "prepositions" and "comparative adjectives" which are used to express relationships between objects. If the models of such relationships can be determined, they help resolve correspondence ambiguities. However, learning models of these relationships requires solving the correspondence problem. We simultaneously learn the visual features defining "nouns" and the differential visual features defining such "binary-relationships" using an EM-based approach.

1 Introduction

There has been recent interest in learning visual classifiers of objects from images with text captions. This involves establishing correspondence between image regions and semantic object classes named by the nouns in the text. There exist significant ambiguities in correspondence of visual features and object classes. For example, figure 1 contains an image which has been annotated with the nouns "car" and "street". It is difficult to determine which regions of the image correspond to which word unless additional images are available containing "street" but not "car" (and vice-versa). A wide range of automatic image annotation approaches use such co-occurrence relationships to address the correspondence problem.

Some words, however, almost always occur in fixed groups, which limits the utility of co-occurrence relationships, alone, to reduce ambiguities in correspondence. For example, since cars are typically found on streets, it is difficult to resolve the correspondence using co-occurrence relationships alone. While such

* The authors would like to thank Kobus Barnard for providing the Corel-5k dataset. The authors would also like to acknowledge VACE for supporting the research.

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part I, LNCS 5302, pp. 16–29, 2008.

confusion is not a serious impediment for image annotation, it is a problem if localization is a goal[1].

We describe how to reduce ambiguities in correspondence by exploiting natural relationships that exists between objects in an image. These relationships correspond to language constructs such as "prepositions" (e.g. above, below) and "comparative adjectives" (e.g. brighter, smaller). If models for such relationships were known and images were annotated with them, then they would constrain the correspondence problem and help resolve ambiguities. For example, in figure 1, consider the binary relationship $on(car, street)$. Using this relationship, we can trivially infer that the green region corresponds to "car" and the magenta region corresponds to "street".

The size of the vocabulary of binary relationships is very small compared to the vocabulary of nouns/objects. Therefore, human knowledge could be tapped to specify rules which can act as classifiers for such relationships (for example, a binary relationship $above(s_1, p_1) \Rightarrow s_1.y < p_1.y$). Alternatively, models can be learned from annotated images. Learning such binary relationships from a weakly-labeled dataset would be "straight forward" if we had a solution to the correspondence problem at hand. This leads to a chicken-egg problem, where models for the binary relationships are needed for solving the correspondence problem, and the solution of the correspondence problem is required for acquiring models of the binary relationships. We utilize an EM-based approach to simultaneously learn visual classifiers of objects and "differential" models of common prepositions and comparative binary relationships.

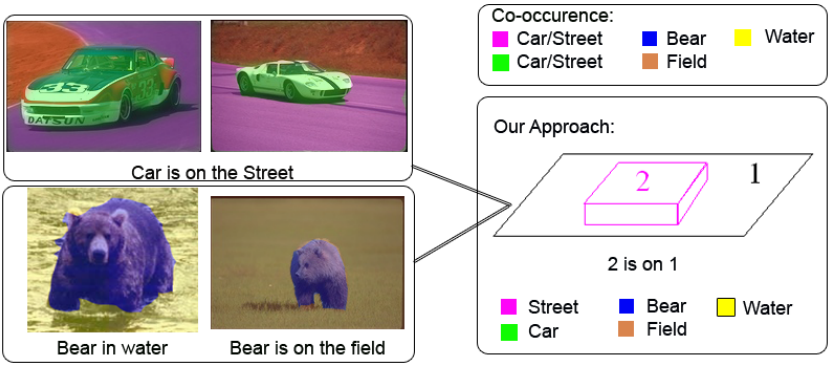


Fig. 1. An example of how our approach can resolve ambiguities. In the case of co-occurrence based approaches, it is hard to correspond the magenta/green regions to 'car'/'street'. 'Bear', 'water' and 'field' are easy to correspond. However, the correct correspondences of 'bear' and 'field' can be used to acquire a model for the relation 'on'. We can then use that model to classify the green region as belonging to 'car' and the magenta one to 'street', since only this assignment satisfies the binary relationship.

[1] It has also been argued [1] that for accurate retrieval, understanding image semantics (spatial localization) is critical.

The significance of the work is threefold: (1) It allows us to learn classifiers (i.e models) for a vocabulary of prepositions and comparative adjectives. These classifiers are based on differential features extracted from pairs of regions in an image. (2) Simultaneous learning of nouns and relationships reduces correspondence ambiguity and leads to better learning performance. (3) Learning priors on relationships that exist between nouns constrains the annotation problem and leads to better labeling and localization performance on the test dataset.

2 Related Work

Our work is clearly related to prior work on relating text captions and image features for automatic image annotation [2,3,4]. Many learning approaches have been used for annotating images which include translation models [5], statistical models [2,6], classification approaches [7,8,9] and relevance language models [10,11].

Classification based approaches build classifiers without solving the correspondence problem. These classifiers are learned on positive and negative examples generated from captions. Relevance language models annotate a test image by finding similar images in the training dataset and using the annotation words shared by them.

Statistical approaches model the joint distribution of nouns and image features. These approaches use co-occurrence counts between nouns and image features to predict the annotation of a test image [12,13]. Barnard et al. [6] presented a generative model for image annotation that induces hierarchical structure from the co-occurrence data. Srikanth et al. [14] proposed an approach to use the hierarchy induced by WordNet for image annotation. Duygulu et al. [5] modeled the problem as a standard machine translation problem. The image is assumed to be a collection of blobs (vocabulary of image features) and the problem becomes analogous to learning a lexicon from aligned bi-text. Other approaches such as [15] also model word to word correlations where prediction of one word induces a prior on prediction of other words.

All these approaches use co-occurrence relationships between nouns and image features; but they cannot, generally, resolve all correspondence ambiguities. They do not utilize other constructs from natural language and speech tagging approaches [16,17]. As a trivial example, given the annotation "pink flower" and a model of the adjective "pink", one would expect a dramatic reduction in the set of regions that would be classified as a flower in such an image. Other language constructs, such as "prepositions" or "comparative adjectives", which express relationships between two or more objects in the image, can also resolve ambiguities.

Our goal is to learn models, in the form of classifiers, for such language constructs. Ferrari et al. [18] presented an approach to learn visual attributes from a training dataset of positive and negative images using a generative model. However, collecting a dataset for all such visual attributes is cumbersome. Ideally we would like to use the original training dataset with captions to learn the appearance of

nouns/adjectives and also understand the meanings of common prepositions and comparative adjectives. Barnard et al. [19] presented an approach for learning adjectives and nouns from the same dataset. They treat adjectives similarly to nouns and use a two step process to learn the models. In the first step, they consider only adjectives as annotated text and learn models for them using a latent model. In the second step, they use the same latent model to learn nouns where learned models of adjectives are used to provide prior probabilities for labeling nouns. While such an approach might be applicable to learning models for adjectives, it cannot be applied to learning models for higher order(binary) relationships unless the models for the nouns are given.

Barnard et al. [20] also presented an approach to reduce correspondence ambiguity in weakly labeled data. They separate the problems of learning models of nouns from resolving correspondence ambiguities. They use a loose model for defining affinities between different regions and use the principal of exclusion reasoning to resolve ambiguities. On the other hand, we propose an approach to simultaneously resolve correspondence ambiguities and learn models of nouns using other language constructs which represent higher order relationships[2].

We also present a systematic approach to employing contextual information (second-order) for labeling images. The use of second order contextual information is very important during labeling because it can help resolve the ambiguities due to appearance confusion in many cases. For example, a blue homogeneous region, B, can be labeled as "water" as well as "sky" due to the similarity in appearance. However, the relation of the region to other nouns such as the "sun" can resolve the ambiguity. If the relation $below(B, sun)$ is more likely than $in(sun, B)$, then the region B can be labeled as "water" (and vice-versa). As compared to [20], which uses adjacency relations for resolution, our approach provides a broader range of relations(prepositions and comparative adjectives) that can be learned simultaneously with the nouns.

3 Overview

Each image in a training set is annotated with nouns and relationships between some subset of pairs of those nouns. We refer to each relationship instance, such as $above(A, B)$, as a predicate. Our goal is to learn classifiers for nouns and relationships (prepositions and comparative adjectives). Similar to [5], we represent each image with a set of image regions. Each image region is represented by a set of visual features based on appearance and shape (e.g area, RGB). The classifiers for nouns are based on these features. The classifiers for relationships are based on differential features extracted from pairs of regions such as the difference in area of two regions.

Learning models of both nouns and relationships requires assigning image regions to annotated nouns. As the data is weakly labeled, there is no explicit assignment of words to image regions. One could, however, assign regions to nouns

[2] The principles of exclusion reasoning are also applicable to our problem. We, however, ignore them here.

if the models of nouns and relationships were known. This leads to a chicken-egg problem. We treat assignment as the missing data and use an EM-approach to learn assignment and models simultaneously. In the E-step we evaluate possible assignments using the parameters obtained at previous iterations. Using the probabilistic distribution of assignment computed in the E-step, we estimate the maximum likelihood parameters of the classifiers in the M-step.

In the next section, we first discuss our model of generating predicates for a pair of image regions. This is followed by a discussion on learning the parameters of the model, which are the parameters of classifiers for nouns, prepositions and comparative adjectives.

4 Our Approach

4.1 Generative Model

We next describe the model for language and image generation for a pair of objects. Figure 2 shows our generative model.

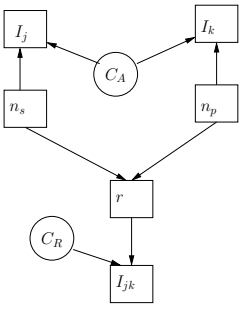


Fig. 2. The Graphical Model for Image Annotation

Each image is represented with a set of image regions and each region is associated with an object which can be classified as belonging to a certain semantic object class. These semantic object classes are represented by nouns in the vocabulary[3].

Assume two regions j and k are associated with objects belonging to semantic object classes, n_s and n_p respectively. Each region is described by a set of visual features I_j and I_k. The likelihood of image features I_j and I_k would depend on

[3] Generally, there will not be a one-one relationship between semantic object classes and nouns. For example, the word "bar" refers to two different semantic concepts in the sentences: "He went to the bar for a drink" and "There were bars in the window to prevent escape". Similarly, one semantic object class can be described by two or more words(synonyms). While dealing with synonyms and word sense disambiguation [21] is an important problem, we simplify the exposition by assuming a one-one relationship between semantic object classes and the nouns in the annotation.

the nouns n_s and n_p and the parameters of the appearance models(C_A) of these nouns. These parameters encode visual appearance of the object classes.

For every pair of image regions, there exist some relationships between them based on their locations and appearances. Relationship types are represented by a vocabulary of prepositions and comparative adjectives. Let r be a type of relationship (such as "above", "below") that holds between the objects associated with regions j and k. The nouns associated with the regions, n_s and n_p, provide priors on the types of relationships in which they might participate (For example, there is a high prior for the relationship "above" if the nouns are "sky" and "water", since in most images "sky" will occur above "water"). Every relationship is described by differential image features I_{jk}. The likelihood of the differential features depends on the type of relationship r and the parameters of the relationship model C_R.

4.2 Learning the Model

The training data consists of images annotated with nouns $(n_1^l, n_2^l..)$ and a set of relationships between these nouns represented by predicates $\mathcal{P}^l$, where l is the image number. Learning the model involves maximizing the likelihood of training images being associated with predicates given in the training data. The maximum likelihood parameters are the parameters of object and relationship classifiers, which are represented by $\theta = (C_A, C_R)$. However, evaluating the likelihood is expensive since it requires summation over all possible assignments of image regions to nouns. We instead treat the assignment as missing data and use an EM formulation to estimate θ^{ML}.

$$\theta^{ML} = \arg\max_{\theta} P(\mathcal{P}^1, \mathcal{P}^2..|I^1, I^2.., \theta) = \arg\max_{\theta} \sum_{A} P(\mathcal{P}^1, \mathcal{P}^2.., A|I^1, I^2.., \theta)$$

$$= \arg\max_{\theta} \prod_{l=1}^{N} \sum_{A^l} P(\mathcal{P}^l|I^l, \theta, A^l) P(A^l|I^l, \theta) \quad (1)$$

where A^l defines the assignment of image regions to annotated nouns in image l. Therefore, $A_i^l = j$ indicates that noun n_i^l is associated to region j in image l.

The first term in equation 1 represents the joint predicate likelihood given the assignments, classifier parameters and image regions. A predicate is represented as

Table 1. Notation

N: Number of images	l: Image under consideration (superscript)
$\mathcal{P}^l$: Set of Predicates for image l	$(n_1^l, n_2^l...)$: Set of Nouns for image l
$\mathcal{P}_i^l = r_i^l(n_{s_i}^l, n_{p_i}^l)$: i^{th} predicate	$A_i^l = j$: Noun n_i^l is associated with region j
C_A: Parameters of models of nouns	C_R:Parameters of models of relationships
r_i^l: Relationship represented by i^{th} predicate	
s_i: Index of noun which appears as argument1 in i^{th} predicate	
p_i: Index of noun which appears as argument2 in i^{th} predicate	
$I_{A_i^l}^l$: Image features for region assigned to noun n_i^l	

$r_i^l(n_{s_i}^l, n_{p_i}^l)$, where r_i^l is a relationship that exists between the nouns associated with region $A_{s_i}^l$ and $A_{p_i}^l$. We assume that each predicate is generated independently of others, given an image and assignment. Therefore, we rewrite the likelihood as:

$$\begin{aligned} P(\mathcal{P}^l|I^l,\theta,A^l) &= \prod_{i=1}^{|\mathcal{P}^l|} P(\mathcal{P}_i^l|I^l,A^l,\theta) \\ &\propto \prod_{i=1}^{|\mathcal{P}^l|} P(r_i^l|I^l_{A^l_{s_i}A^l_{p_i}},C_R)P(r_i^l|n_{s_i},n_{p_i}) \\ &\propto \prod_{i=1}^{|\mathcal{P}^l|} P(I^l_{A^l_{s_i}A^l_{p_i}}|r_i^l,C_R)P(r_i^l|C_R)P(r_i^l|n_{s_i},n_{p_i}) \end{aligned}$$

Given the assignments, the probability of associating a predicate $\mathcal{P}_i^l$ to the image is the probability of associating the relationship r_i^l to the differential features associated with the pair of regions assigned to n_{s_i} and n_{p_i}. Using Bayes rule, we transform this into the differential feature likelihood given the relationship word and the parameters of the classifier for that relationship word. $P(r_i^l|C_R)$ represents the prior on relationship words and is assumed uniform.

The second term in equation 1 evaluates the probability of an assignment of image regions to nouns given the image and the classifier parameters. Using Bayes rule, we rewrite this as:

$$\begin{aligned} P(A^l|I^l,\theta) &= \prod_{i=1}^{|A^l|} P(n_i^l|I^l_{A_i^l},C_A) \\ &\propto \prod_{i=1}^{|A^l|} P(I^l_{A_i^l}|n_i^l,C_A)P(n_i^l|C_A) \end{aligned}$$

where $|A^l|$ is the number of annotated nouns in the image, $P(I^l_{A_i^l}|n_i^l,C_A)$ is the image likelihood of the region assigned to the noun, given the noun and the parameters of the object model, $P(n_i^l|C_A)$ is the prior over nouns given the parameters of object models.

EM-Approach. We use an EM approach to simultaneously solve for the correspondence and for learning the parameters of classifiers represented by θ.

1. **E-step:** Compute the noun assignment for a given set of parameters from the previous iteration represented by θ^{old}. The probability of assignment in which noun i correspond to region j is given by:

$$P(A_i^l = j|\mathcal{P}^l,I^l,\theta^{old}) = \frac{\sum_{A'\in\mathcal{A}_{ij}^l} P(A'|\mathcal{P}^l,I^l,\theta^{old})}{\sum_k \sum_{A'\in\mathcal{A}_{ik}^l} P(A'|\mathcal{P}^l,I^l,\theta^{old})} \tag{2}$$

where $\mathcal{A}_{ij}$ refers to the subset of the set of all possible assignments for an image in which noun i is assigned to region j. The probability of any assignment A' for the image can be computed using Bayes rule:

$$P(A'|\mathcal{P}^l,I^l,\theta^{old}) \propto P(\mathcal{P}^l|A',I^l,\theta^{old})P(A'|I^l,\theta^{old}) \tag{3}$$

2. **M-step:** For the noun assignment computed in the E-step, we find the new ML parameters by learning both relationship and object classifiers. The ML parameters depend on the type of classifier used. For example, for a gaussian classifier we estimate the mean and variance for each object class and relationship class.

For initialization of the EM approach, we can use any image annotation approach with localization such as the translation based model described in [5]. Based on initial assignments, we initialize the parameters of both relationship and object classifiers.

We also want to learn the priors on relationship types given the nouns represented by $P(r|n_s, n_p)$. After learning the maximum likelihood parameters, we use the relationship classifier and the assignment to find possible relationships between all pairs of words. Using these generated relationship annotations we form a co-occurrence table which is used to compute $P(r|n_s, n_p)$.

4.3 Inference

Similar to training, we first divide the test image into regions. Each region j is associated with some features I_j and noun n_j. In this case, I_j acts as an observed variable and we have to estimate n_j. Previous approaches estimate nouns for regions independently of each other. We want to use priors on relationships between pair of nouns to constrain the labeling problem. Therefore, the assignment of labels cannot be done independently of each other. Searching the space of all possible assignments is infeasible.

We use a Bayesian network to represent our labeling problem and use belief propagation for inference. For each region, we have two nodes corresponding to the noun and image features from that region. For all possible pairs of regions, we have another two nodes representing a relationship word and differential features

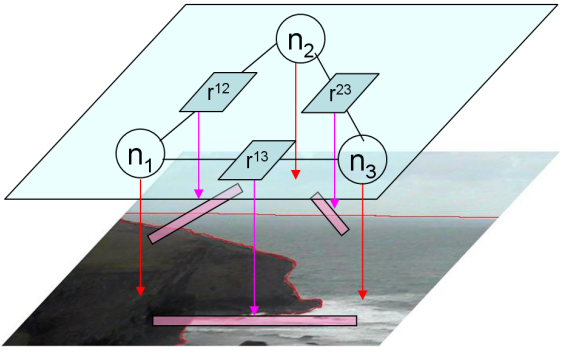


Fig. 3. An example of a Bayesian network with 3 regions. The r^{jk} represent the possible words for the relationship between regions (j, k). Due to the non-symmetric nature of relationships we consider both (j, k) and (k, j) pairs (in the figure only one is shown). The magenta blocks in the image represent differential features (I_{jk}).

from that pair of regions. Figure 3 shows an example of an image with three regions and its associated Bayesian network. The word likelihood is given by:

$$P(n_1, n_2..|I_1, I_2..I_{12}, .., C_A, C_R) \propto \prod_i P(I_i|n_i, C_A) \prod_{(j,k)} \sum_{r_{jk}} P(I_{jk}|r_{jk}, C_R) P(r_{jk}|n_j, n_k) \quad (4)$$

5 Experimental Results

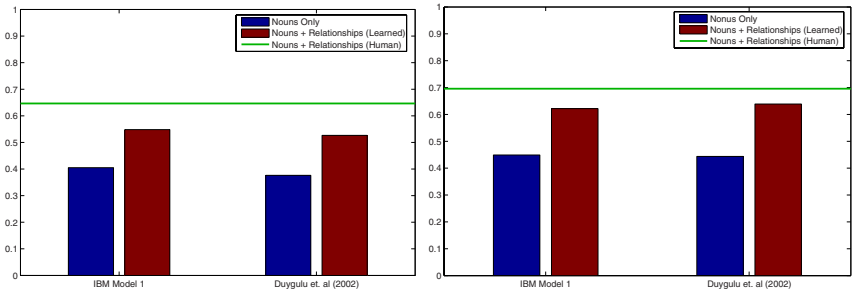
In all the experiments, we use a nearest neighbor based likelihood model for nouns and decision stump based likelihood model for relationships. We assume each relationship model is based on one differential feature(for example, the relationship "above" is based on difference in y locations of 2 regions). The parameter learning M-step therefore also involves feature selection for relationship classifiers. For evaluation we use a subset of the Corel5k training and test dataset used in [5]. For training we use 850 images with nouns and hand-labeled the relationships between subsets of pairs of those nouns. We use a vocabulary of 173 nouns and 19 relationships[4].

5.1 Resolution of Correspondence Ambiguities

We first evaluate the performance of our approach for the resolution of correspondence ambiguities in the training dataset. To evaluate the localization performance, we randomly sampled 150 images from the training dataset and compare it to human labeling. Similar to [22], we evaluate the performance in terms of two measures: "range of semantics identified" and "frequency correct". The first measure counts the number of words that are labeled properly by the algorithm. In this case, each word has similar importance regardless of the frequency with which it occurs. In the second case, a word which occurs more frequently is given higher importance. For example, suppose there are two algorithms one of which only labels 'car' properly and other which only labels 'sky' properly. Using the first measure, both algorithms have similar performance because they can correctly label one word each. However, using the second measure the latter algorithm is better as sky is more common and hence the number of correctly identified regions would be higher for the latter algorithm.

We compare our approach to image annotation algorithms which can be used for localization of nouns as well. These approaches are used to bootstrap our EM-algorithm. For our experiments, a co-occurrence based translation model [13] and translation based model with mixing probabilities [5] form the baseline algorithms. To show the importance of using "prepositions" and "comparative adjectives" for resolution of correspondence ambiguities, we use both algorithms to bootstrap EM and present our results. We also compare our performance with the algorithm where relationships are defined by a human instead of learning them from the dataset itself. Figure 4 compares the performance of all the

[4] Above, behind, below, beside, more textured, brighter, in, greener, larger, left, near, far from, ontopof, more blue, right, similar, smaller, taller, shorter.



(a) Semantic Range (b) Frequency Correct

Fig. 4. Comparison of normalized "semantic range" and "frequency correct" scores for the training dataset. The performance increases substantially by using prepositions and comparative adjectives in addition to nouns. The green line shows the performance when relationships are not learned but are defined by a human. The two red blocks show the performance of our approach where relationships and nouns are learned using the EM algorithm and bootstrapped by IBM Model1 or Duygulu et al. respectively.

algorithms with respect to the two measures described above. Figure 5 shows some examples of how ambiguity is removed using prepositions and comparative adjectives.

5.2 Labeling New Images

We also tested our model on labeling new test images. We used a subset of 500 test images provided in the Corel5k dataset. The subset was chosen based on the vocabulary of nouns learned from the training. The images were selected randomly from those images which had been annotated with the words present in our learned vocabulary. To find the missed labels we compute $\mathcal{S}_t \setminus \mathcal{S}_g$, where $\mathcal{S}_t$ is the set of annotations provided from the Corel dataset and $\mathcal{S}_g$ is the set of annotations generated by the algorithm. However, to test the correctness of labels generated by the algorithm we ask human observers to verify the annotations. We do not use the annotations in the Corel dataset since they contain only a subset of all possible nouns that describe an image. Using Corel annotations for evaluation can be misleading, for example, if there is "sky" in an image and an algorithm generates an annotation "sky" it may be labeled as incorrect because of the absence of sky from the Corel annotations. Figure 6 shows the performance of the algorithm on the test dataset. Using the proposed Bayesian model, the number of missed labels decreases by 24% for IBM Model 1 and by 17% for Duygulu et al. [5]. Also, using our approach 63% and 59% of false labels are removed respectively.

Figure 7 shows some examples of the labeling on the test set. The examples show how Bayesian reasoning leads to better labeling by applying priors on relationships between nouns. The recall and precision ratios for some common words in the vocabulary are shown in Table 2. The recall ratio of a word represents

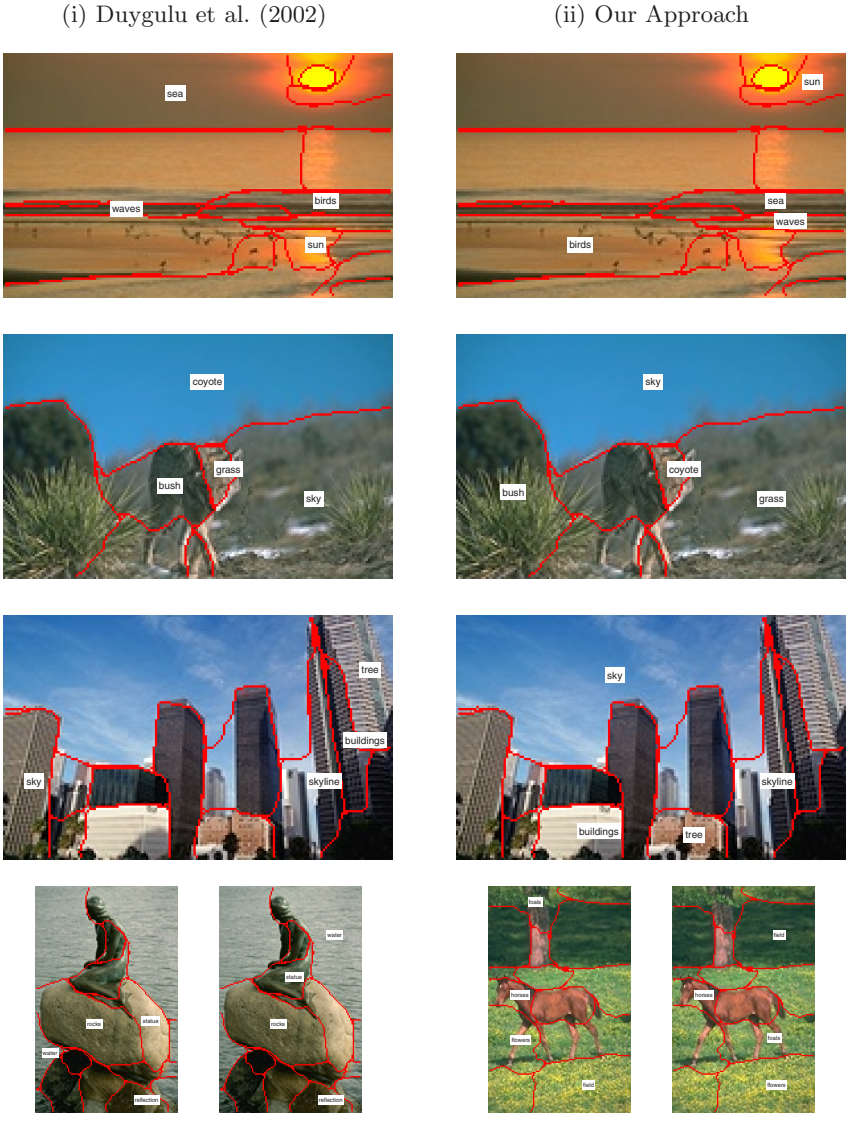
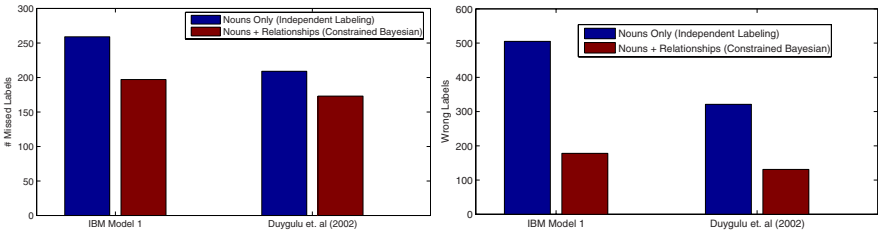


Fig. 5. Some examples of how correspondence ambiguity can be reduced using prepositions and comparative adjectives. Some of the annotations for the images are: **(a)** *near(birds,sea); below(birds,sun); above(sun, sea); larger(sea,sun); brighter(sun, sea); below(waves,sun)* **(b)** *below(coyote, sky); below(bush, sky); left(bush, coyote); greener(grass, coyote); below(grass,sky)* **(c)** *below(building, sky); below(tree,building); below(tree, skyline); behind(buildings,tree) blueish(sky, tree)* **(d)** *above(statue,rocks); ontopof(rocks, water); larger(water,statue)* **(e)** *below(flowers,horses); ontopof(horses, field); below(flowers,foals).*



(a) Missed Labels (b) False Labels

Fig. 6. Labeling performance on set of 100 test images. We do not consider localization errors in this evaluation. Each image has on average 4 labels in the Corel dataset.

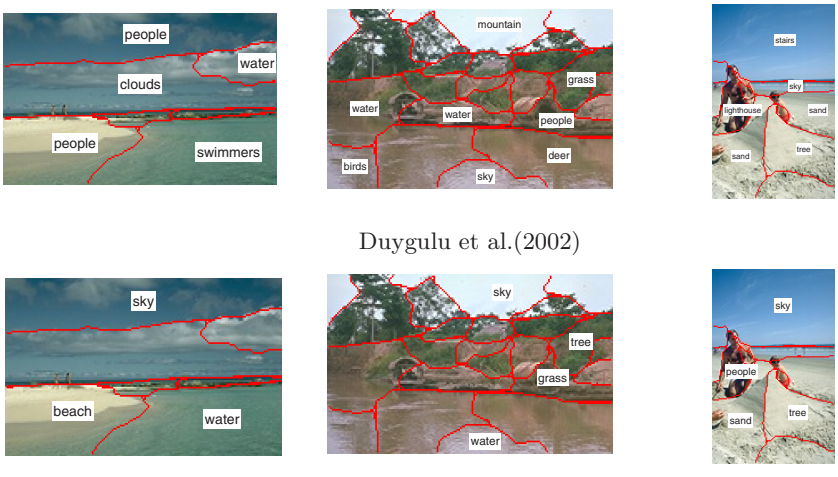


Fig. 7. Some examples of labeling on test dataset. By applying priors on relationships between different nouns, we can improve the labeling performance. For example, when labels are predicted independently, there can be labeling where region labeled "water" is above region labeled "clouds" as shown in the first image. This is however incongruent with the priors learned from training data where "clouds" are mostly above "water". Bayesian reasoning over such priors and likelihoods lead to better labeling performance.

the ratio of the number of images correctly annotated with that word using the algorithm to the number of images that should have been annotated with that word. The precision ratio of a word is the ratio of number of images that have been correctly annotated with that word to the number of images which were annotated with the word by the algorithm. While recall rates are reported with respect to corel annotations, precision rates are reported with respect to correctness defined by human observers. The results show that using a constrained bayesian model leads to improvement in labeling performance of common words in terms of both recall and precision rates.

Table 2. Precision-Recall Ratios

	water	grass	clouds	buildings	sun	sky	tree	
Recall	0.79	0.7	0.27	0.25	0.57	0.6	0.66	Duygulu(2002)
	0.90	1.0	0.27	0.42	0.57	0.93	0.75	Ours
Precision	0.57	0.84	0.76	0.68	0.77	0.98	0.70	Duygulu(2002)
	0.67	0.79	0.88	0.80	1.00	1.00	0.75	Ours

6 Conclusion

Learning visual classifiers from weakly labeled data is a hard problem which involves finding a correspondence between the nouns and image regions. While most approaches use a "bag" of nouns model and try to find correspondence using co-occurrence of image features and the nouns, correspondence ambiguity remains. We proposed the use of language constructs other than nouns, such as prepositions and comparative adjectives, to reduce correspondence ambiguity. While these relationships can be defined by humans, we present an EM based approach to simultaneously learn visual classifiers for nouns, prepositions and comparative adjectives. We also present a more constrained Bayesian model for the labeling process. Experimental results show that using relationship words helps in reduction of correspondence ambiguity and using a constrained model leads to a better labeling performance.

References

1. Armitage, L., Enser, P.: Analysis of user need in image archives. Journal of Information Science (1997)
2. Barnard, K., Duygulu, P., Freitas, N., Forsyth, D., Blei, D., Jordan, M.I.: Matching words and pictures. Journal of Machine Learning Research, 1107–1135 (2003)
3. Carneiro, G., Chan, A.B., Moreno, P., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE PAMI (2007)
4. Carbonetto, P., Freitas, N., Barnard, K.: A statistical model for general contextual object recognition. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 350–362. Springer, Heidelberg (2004)
5. Duygulu, P., Barnard, K., Freitas, N., Forsyth, D.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2353, pp. 97–112. Springer, Heidelberg (2002)
6. Barnard, K., Forsyth, D.: Learning the semantics of words and pictures. In: ICCV, pp. 408–415 (2001)
7. Andrews, S., Tsochantaridis, I., Hoffman, T.: Support vector machines for multiple-instance learning. In: NIPS (2002)
8. Li, J., Wang, J.: Automatic linguistic indexing of pictures by statistical modeling approach. IEEE PAMI (2003)
9. Maron, O., Ratan, A.: Multiple-instance learning for natural scene classification. ICML (1998)
10. Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: NIPS (2003)

11. Feng, S., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. In: CVPR (2004)
12. Mori, Y., Takahashi, H., Oka, R.: Image to word transformation based on dividing and vector quantizing images with words. MISRM (1999)
13. Brown, P., Pietra, S., Pietra, V., Mercer, R.: The mathematics of statistical machine translation: Parameter estimation. Computational Linguistics (1993)
14. Srikanth, M., Varner, J., Bowden, M., Moldovan, D.: Exploiting ontologies for automatic image annotation. SIGIR (2005)
15. Jin, R., Chai, J., Si, L.: Effective automatic image annotation via a coherent language model and active learning. Mutimedia (2004)
16. Brill, E.: A simple rule-based part of speech tagger. ACL (1992)
17. Brill, E.: Transformation-based error-driven learning and natural language processing. Computational Linguistics (1995)
18. Ferrari, V., Zisserman, A.: Learning visual attributes. In: NIPS (2007)
19. Barnard, K., Yanai, K., Johnson, M., Gabbur, P.: Cross modal disambiguation. Toward Category-Level Object Recognition (2006)
20. Barnard, K., Fan, Q.: Reducing correspondence ambiguity in loosely labeled training data. In: CVPR (2007)
21. Barnard, K., Johnson, M.: Word sense disambigutaion with pictures. AI (2005)
22. Barnard, K., Fan, Q., Swaminathan, R., Hoogs, A., Collins, R., Rondot, P., Kaufold, J.: Evaluation of localized semantics: data, methodology and experiments. Univ. of Arizona, TR-2005 (2005)