Detecting Objects via Multiple Segmentations and Latent Topic Models [*]

Tomasz J. Malisiewicz
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
tmalisie@cs.cmu.edu

Jonathan C. Huang
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
jch1@cs.cmu.edu

Alexei A. Efros
The Robotics Institute
Carnegie Mellon University
Pittsburgh, PA 15213
efros@cs.cmu.edu

Abstract

We present an object detection system and show results on the PASCAL 2006 dataset. We introduce a codebook defined over segment-level features and show how multiple segmentations and Latent Topic Models can be used to localize objects in images despite the bag-of-words assumption. We demonstrate how to train both the Latent Dirichlet Allocation model and the Correlated Topic Model in a supervised way and show how our approach is capable of detecting multiple instances of objects in images such as cars, horses, grass, sheep, etc.

1 Introduction

Object detection can be defined as the localization of novel object instances in an image from a known list of object classes. This problem is difficult because of the appearance, geometry, and viewpoint variations that can occur in an image. To make things worse, not only are many object classes visually similar but we want to avoid training specialized detectors for each different object class.

Given a manually labeled dataset that contains either bounding boxes or segmentation masks of objects from a small number of object classes, we want to learn models of those objects so that we can localize novel instances of those objects in a new image.

1.1 What has been done before?

In the recent years, several different approaches have been presented with the goal of detecting objects in images using a codebook of appearance parts.

[*] This report is focused on the Computer Vision aspect of this work, while the Correlated Topic Model Details report is focused on the Machine Learning aspect of this work.

While Fergus *et al.* [6] utilized a strongly parameterized geometric model over parts and Vidal-Naquet & Ullman [13] used loose relationships between parts, bag-of-words approaches which don't model any geometrical relationships between parts have perform suprisingly well. Examples of bag-of-words approaches include the work of Csurka *et al.* [4] and Sivic *et al.* [12].

In addition, a large body of work has focused on incorporating context into object detection. Much of this work is based on some form of a conditional random field (Lafferty *et al.* [10]). CRF approaches avoid making any hard decisions locally by employing both local classifiers and pairwise relationships between these classifiers. Statistical inference algorithms are then used to find the best assignment of image features to classes.

2 Representation

2.1 Bag-of-words

Building off of the successes of the bag-of-words approach, we also represent an image as a bag of words; however, we define a lexicon over segment-level words (s-words) which densely cover the image. By utilizing the bag-of-words representation, we discard all information regarding the connectivity of segments. These s-words are the most atomic units of representation in our work; the final segmentation mask obtained by our algorithm is a linear combination of these atomic s-words.

2.2 Multiple Segmentations

Although using segments allows us to densely cover the image (as opposed to sparse visual words [12]), we are still left with the problem of obtaining a good segmentation. Instead of trusting a single bottom-up segmentation to produce a correct grouping of pixels, we use a soup of segments (several outputs of a segmentation algorithm by varying its parameters). We hope that by using multiple segmentations, some of the segments will be correct some of the time. An example of multiple segmentations is show in Figure 2 for the cow image found in Figure 1.

Multiple segmentations have been used in the work of Russel *et al.* [11] and Hoiem *et al.* [8] for slightly different purposes. Our goal is to detect objects in an image while Russell *et al.* [11] used multiple segmentations for the unsupervised discovery of semantically meaningful segments in a large collection of images and Hoiem *et al.* [8] used multiple segmentations to classify image regions into one of several geometric classes.

2.3 Texton Histograms

After obtaining multiple segmentations for each training image, we describe each segment by a normalized histogram of textons. Our textons are obtained by vector quantizing filter bank responses via K-means to T clusters. We use a filter bank very similar to that of Winn *et al.* [14], which contains both zero-sum and sum-to-one filters that are convolved with the 3 LAB channels of an image. In other words, certain filters respond to color while other filters respond to edges at different orientations. Each pixel in the training images is then assigned to the nearest texton.

Next, the T-component texton histograms are normalized to achieve robustness to scale variations and then vector quantized via K-means to L clusters. (In the language of statistical text processing, we would say that we have L words in our lexicon.) Each segment-level normalized texton histogram is then assigned to the nearest cluster center; these L s-words are the primitives that we use as input into the Bayesian Hierarchical Models defined in the following section.

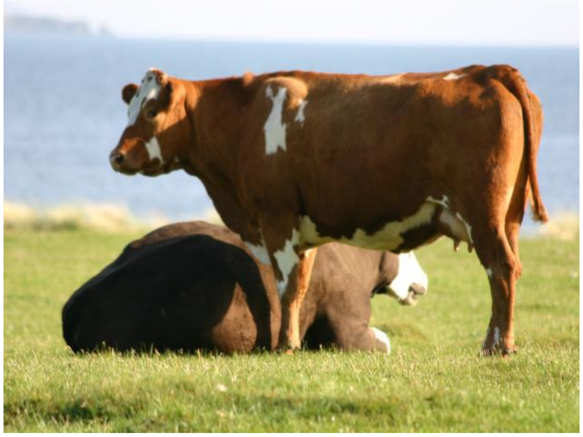

Figure 1: Running example cow image from PASCAL 2006 dataset.

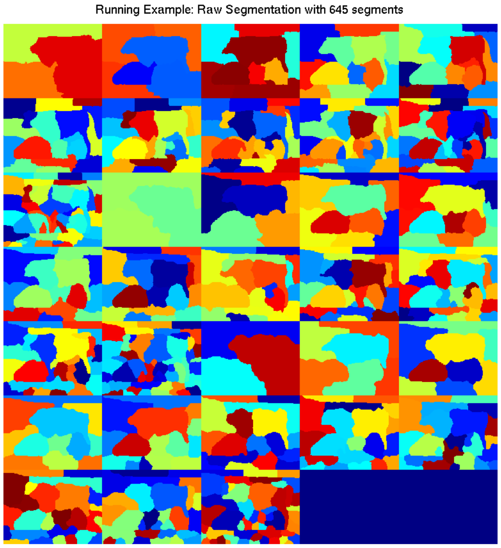

Figure 2: Segments extracted from a PASCAL 2006 image. (33 Segmentations)

3 Latent Topic Models

Latent Topic Models (otherwise known as aspect models) have been successfully used for unsupervised topic discovery in text corpora [3]. The strength of these techniques is their ability to deal with polysemous words. For example, the word *bank* could refer to a *river bank* or a *financial bank*. Based on co-occurrence of words from a finite lexicon, topic model are able to associate words a topic. These techniques have also been used for scene classification by Fei-Fei *et al.* [5] and object discovery by Sivic *et al.* [12]. Fei-Fei *et al.* [5] used a codebook from image patches, while Sivic *et al.* [12] used a codebook made from SIFT descriptors computed at interest points.

3.1 Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) model by Blei et al [3] is a generative model for a collection of exchangeable discrete data. LDA has mainly been used to model text corpora, where the notion of exchangeability corresponds to the bag-of-words assumption that is commonly employed in such models.

The model is simplest to describe for text corpora. In a nutshell, LDA models each document as a mixture over topics, where each vector of mixture proportions is assumed to have been drawn from a Dirichlet distribution. A topic in this model is defined to be a discrete distribution over words from some finite lexicon. For example, if a topic is “astrophysics”, then the word “quasar” would presumably be assigned a higher probability than the word “burrito”.

More precisely, let D be a collection of M documents and $\mathbf{w}$ be a document with words $w_1, \ldots, w_n$. We assume that each word corresponds to one of K possible topics, and that for each word, there is a latent topic assignment z_i which takes on values $1, \ldots, K$, indexing into the set of topics. By convention, we will represent z_i by a K dimensional vector with one component set to one indicating its value in $1, \ldots, K$ and the rest set to zero. The generative process is as follows:

1. Draw $\theta \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_k)$
2. For each word $w_n \in \mathbf{w}$,
    (a) Draw a topic $z_i \sim$ Multinomial(θ)
    (b) Draw a word $w_i \sim$ Multinomial$(\beta(z_i))$ where $\beta(k)$ is a probability distribution over words corresponding to topic k

The inference task in LDA is to solve for the vector of topic mixture proportions, θ, and the topic assignment z_i for each word, given the words w_i and the model parameters α, β_k. Exact inference in this model involves an intractable integral, but approximations via Markov Chain Monte Carlo sampling [7] or mean-field methods have been shown to work well.

3.2 Correlated Topic Model

A problem with drawing the topic mixture proportions (θs) from a Dirichlet distribution is that Dirichlet distributions are too simple, and exhibit a near-independence structure.[1] However, visual object classes are highly correlated and thus we expect the Dirichlet distribution to lack the power to model the correlations between objects. For example bird/sky, sheep/grass, car/road are highly correlated pairs of object classes.

[1] A way to draw from a Dirichlet distribution is to sample k independent Gamma distributions, concatenate the samples into a vector and divide by their sum. This process shows that the correlations between the components of a Dirichlet random variable arise solely from the fact that they must all sum to one.

The solution is to model each mixture over topics as a sample from a more powerful distribution. The *Logistic-Normal* distribution [1] is an alternative distribution over a simplex which forms a richer class of distributions and better captures inter-component correlations. The process of drawing a k-dimensional Logistic-Normal random variable θ is as follows:

1. Draw $v \sim N(\mu, \Sigma)$ where $N(\mu, \Sigma)$ is a $k-1$ dimensional Normal distribution.
2. Define $v_k = 0$.
3. Let
$$\theta = \frac{\exp v}{\sum_{j=1}^{k} \exp v_j}$$
(This is the projection of $\exp(v)$ to the simplex)

We are thus able to capture the covariance between visual classes using the *Logistic-Normal* at the expense of computational complexity. The Correlated Topic Model [2] models the same type of data as LDA and only differs in the first step of the generative process. Instead of drawing θ from a Dirichlet distribution it assumes that θ is drawn from a Logistic-Normal Distribution.

3.3 Supervised Training

Both LDA and CTM have been introduced for the unsupervised discovery of topics in text corpora. For our particular application, we utilize ground truth labels to supervise the training procedure. Supervised training for Latent Topic Models allows us to define what is a topic. In our work, we enforce a one-to-one correspondence between topics and object classes. It is important to note that by supervising our training procedure we are still using the full LDA/CTM models; the only difference is that during training the topics are observed. Interestingly, since our framework is based on models that have been shown to work in an unsupervised setting we can perform training in an unsupervised way with minimal modifications to our approach (we choose not to pursue this direction at the moment).

3.4 Background Modeling & Segment Labeling

In order to perform supervised training, we need a topic label for each segment in the training set. Initial experiments on the MSRC dataset [2], which contains ground truth segmentation masks for background categories such as grass, road, sky, and water demonstrated the utility of modeling background categories as well as objects such as cats, sheep, cars, etc. Unfortunately, many object datasets only contain ground truth labels for objects and the backgrounds are unlabeled. For such datasets, we create pseudo-background labels by clustering background segments on appearance. When labeling the segments from the training set with ground truth labels, we employed the following procedure:

Segments falling completely inside an object of interest were given the label of that particular class. Segments falling completely outside of any objects, were labelled as “background”. Segments which contained pixels both inside a bounding box and in the background were labelled as the foreground object if their overlap score was better than O_{good}, labelled as background if their overlap was less than O_{bg} and labelled as “bad” otherwise. (These numbers were tweaked by hand on a small number of training images).

All of the “background” segments from the training images were then clustered into K_{bg} clusters using K-means. Like before, we used the normalized texture histogram as the feature associated with each segment. Finally, each “background” segment was assigned

[2] Publicly available at *http://research.microsoft.com/vision/cambridge/recognition/*

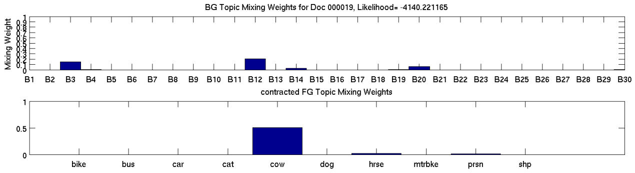

Figure 3: Topic Distribution from LDA inference on a PASCAL 2006 image.

to the nearest cluster. In conclusion the training set consisted of a collection of images where each image contained a bag of segment words with their associated labels. All of (non-"bad") segments were used for training; each segment had one of K labels (K_{bg} background labels and K_{fg} object labels).

3.5 Inference

What can a Latent Topic Model tell us about a new image? For approximate inference, we use variational inference ([9], [3]) which returns an approximate topic distribution per image as well as a topic distribution per word. For example, the topic distribution for the cow image in Figure 1 can be seen in Figure 3.

4 Integrating Multiple Segmentations

We obtain a distribution over topic labels per segment in the soup of segments associated with each image; however, these segments are overlapping and so we must somehow integrate these multiple hypothesis into a single resulting image. Similar to Hoiem *et al.* [8], we marginalize over the multiple segmentations to obtain a topic distribution at the pixel level; for each topic we obtain a topic response images whose pixel values range from 0 to 1. What this marginalization procedure does is simply take the average topic response in a pixel from all segments that contain that particular pixel. A marginalization of the PASCAL 2006 cow image can be seen in Figure 4.

4.1 From Topic Responses to Detection and Segmentation

After we marginalize across segmentations to obtain topic response images for each topic, we can assign each pixel to the most likely topic. This assignment of pixel to topic produces a segmentation and a detection simultaneously. We deem each connected component in the final segmentation corresponding to an object classes as a separate detection. The confidence associated with each such segment is the average topic response inside the segment.

5 Results

5.1 PASCAL Dataset

The dataset we are working with is a subset of the 2006 PASCAL Visual Object Classes Challenge dataset([3]). The subset (referred to as *trainval*) contains $2,618$ images of 10 object classes (bicycle, bus, car, cat, cow, dog, horse, motorbike, person, sheep) with ground-truth bounding boxes at the object-instance level. This dataset is further partitioned into a

[3] http://www.pascal-network.org/challenges/VOC/voc2006/index.html

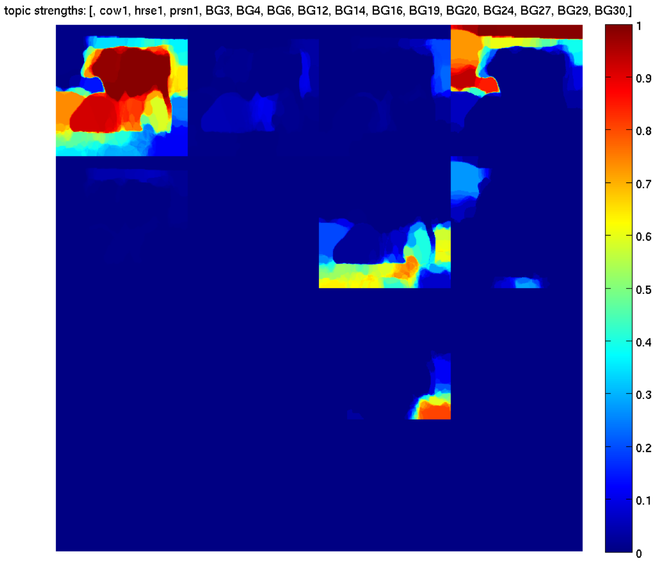

Figure 4: Topic Response Images extracted from a PASCAL 2006 image. The first three sub-images belong to the object classes car, horse, and person while the remaining sub-images correspond to background topics. Not how the first sub-image correctly depicts a strong response to the cow topic.

training set (referred to as *train* in the challenge) and a testing set (referred to as *val* in the challenge). For this subset of the PASCAL dataset, ground truth data was made publicly available which we used for training, testing, and generating precision-recall curves.

Since the PASCAL 2006 dataset data did not contain any annotation for the backgrounds, we created pseudo-background labels by clustering background segments (See Section 3.4).

In order to make our output adhere to the PASCAL challenge, we simply draw a bounding box around each connected component in the resulting marginalized topic response image (see Figure 4).

5.2 Implementation Details

We used a texton dictionary of size $T = 1000$, and an s-word lexicon of size $L = 1500$ in our implementation. Since our training set consisted of 1277 images, random subset of all of features were used for both the texton dictionary and s-word lexicon.

For the overlap criterion, we used $O_{good} = .95$, $O_{bg} = .3$. For the number of background topics, we used $K_{bg} = 30$ and $K_{fg} = 10$.

5.3 Getting Multiple Segmentations

We use Stella Yu's Constrained Normalized Cuts framework [15], and produce 33 segmentations by varying the number of segments (from 3 to 30 by 3s as well as 50 segments) for a total of 11 segmentations per scale and processing the image at three different scales (height of 400, 200, 150 and width changed accordingly). Each image was composed of approximately 650 segments (note that segments from different segmentations are usually

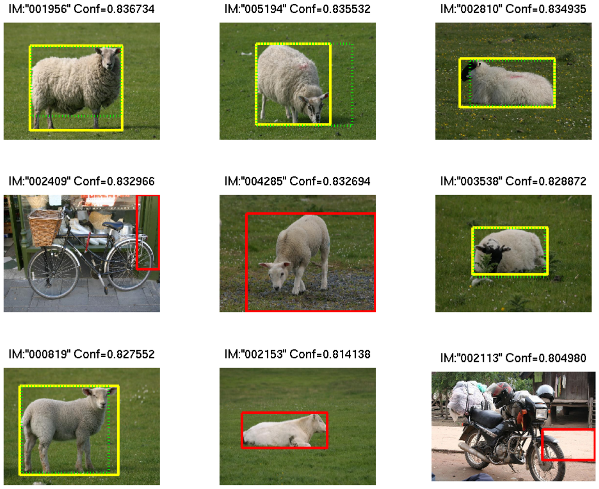
IM:"001956" Conf=0.836734
IM:"005194" Conf=0.835532
IM:"002810" Conf=0.834935
IM:"002409" Conf=0.832966
IM:"004285" Conf=0.832694
IM:"003538" Conf=0.828872
IM:"000819" Conf=0.827552
IM:"002153" Conf=0.814138
IM:"002113" Conf=0.804980

Figure 5: Top 9 sheep detections for the Correlated Topic Model.

overlapping).

5.4 Detection Results

For several object categories and both LDA and CTM we show the top 9 object detections (sorted by confidence). In Figure 5 and Figure 6 we show the top sheep detections. In Figure 7 and Figure 8 we show the top cow detections. In Figure 9 and Figure 10 we show the top bicycle detections. In each figure, a dotted green bounding box represents a true positive, a red bounding box represents a false positive, and a yellow bounding box represents the ground truth bounding box.

5.5 Precision-Recall Curves

In figures 11 and 12 we show the precision-recall curves for the Latent Dirichlet Allocation and Correlated Topic Models, respectively. At the top right hand corner of these figures, the average precision is listed for each object class.

6 Discussion and Summary

6.1 Bag-of-segments and LDA Discussion

Bayesian Hierarchical Models used in this work, namely LDA and CTM, model the context between co-occuring s-words. Independently, each s-word isn't powerful enough to predict the presence of an object of interest; however, a large number of co-occuring s-words are capable of predicting the presence of an object. In our work, the s-word representation was learned in an unsupervised way using k-means clustering and we were able to obtain a good assignment of s-word to topic based on the co-occurrence of s-words in an image.

While we decided to learn appearance clusters in an unsupervised way and obtain a topic distribution per segment for all segments at once, many approaches (especially the CRF-

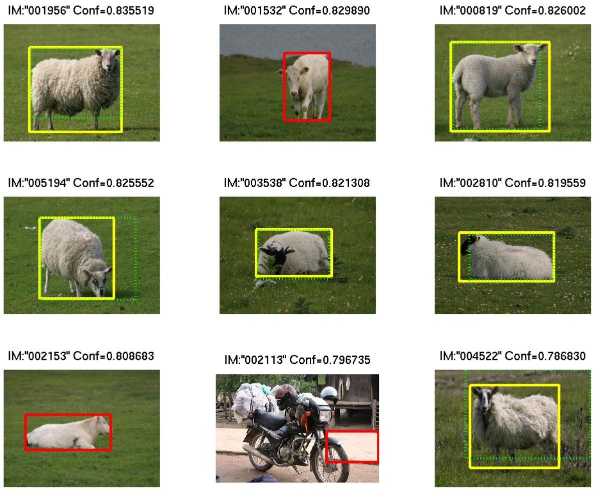
IM:"001956" Conf=0.835519
IM:"001532" Conf=0.829890
IM:"000819" Conf=0.826002
IM:"005194" Conf=0.825552
IM:"003538" Conf=0.821308
IM:"002810" Conf=0.819559
IM:"002153" Conf=0.808683
IM:"002113" Conf=0.796735
IM:"004522" Conf=0.786830

Figure 6: Top 9 sheep detections for the LDA Model.

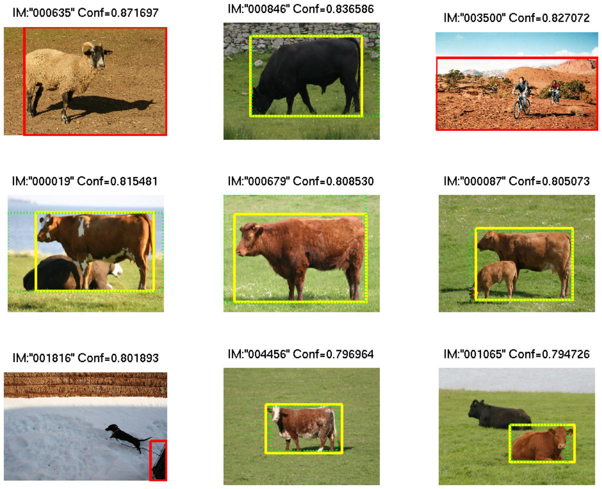
IM:"000635" Conf=0.871697
IM:"000846" Conf=0.836586
IM:"003500" Conf=0.827072
IM:"000019" Conf=0.815481
IM:"000679" Conf=0.808530
IM:"000087" Conf=0.805073
IM:"001816" Conf=0.801893
IM:"004456" Conf=0.796964
IM:"001065" Conf=0.794726

Figure 7: Top 9 cow detections for the Correlated Topic Model.

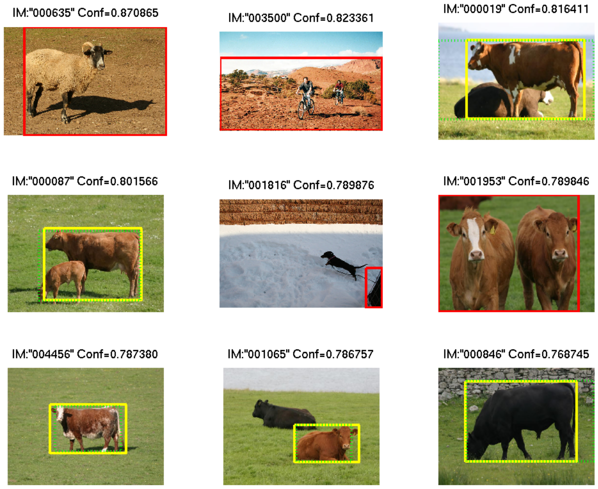
IM:"000635" Conf=0.870865
IM:"003500" Conf=0.823361
IM:"000019" Conf=0.816411
IM:"000087" Conf=0.801566
IM:"001816" Conf=0.789876
IM:"001953" Conf=0.789846
IM:"004456" Conf=0.787380
IM:"001065" Conf=0.786757
IM:"000846" Conf=0.768745

Figure 8: Top 9 cow detections for the LDA Model.

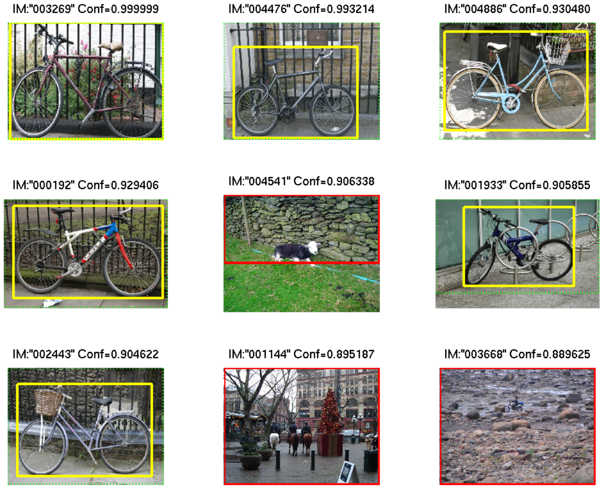
IM:"003269" Conf=0.999999
IM:"004476" Conf=0.993214
IM:"004886" Conf=0.930480
IM:"000192" Conf=0.929406
IM:"004541" Conf=0.906338
IM:"001933" Conf=0.905855
IM:"002443" Conf=0.904622
IM:"001144" Conf=0.895187
IM:"003668" Conf=0.889625

Figure 9: Top 9 bicycle detections for the Correlated Topic Model.

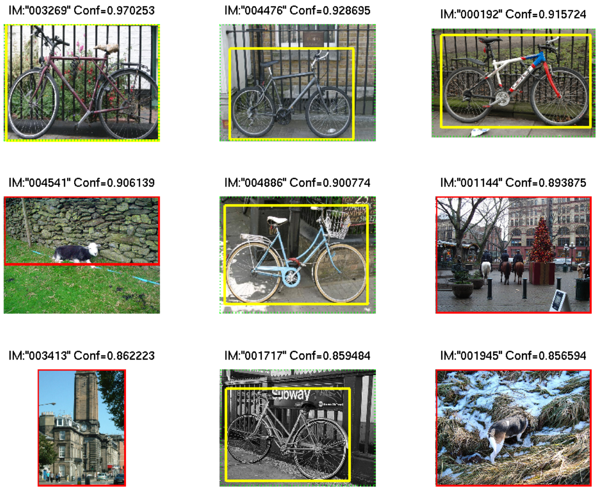
IM:"003269" Conf=0.970253
IM:"004476" Conf=0.928695
IM:"000192" Conf=0.915724
IM:"004541" Conf=0.906139
IM:"004886" Conf=0.900774
IM:"001144" Conf=0.893875
IM:"003413" Conf=0.862223
IM:"001717" Conf=0.859484
Subway
IM:"001945" Conf=0.856594

Figure 10: Top 9 bicycle detections for the LDA Model.

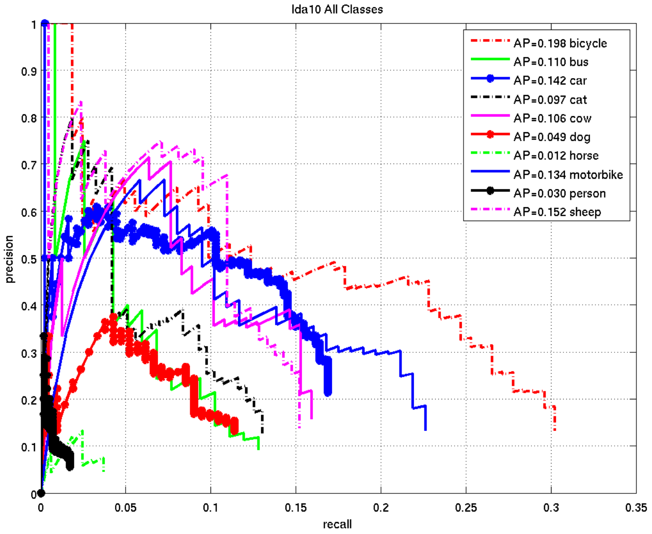

Figure 11: LDA precision-recall curves and Average Precision.

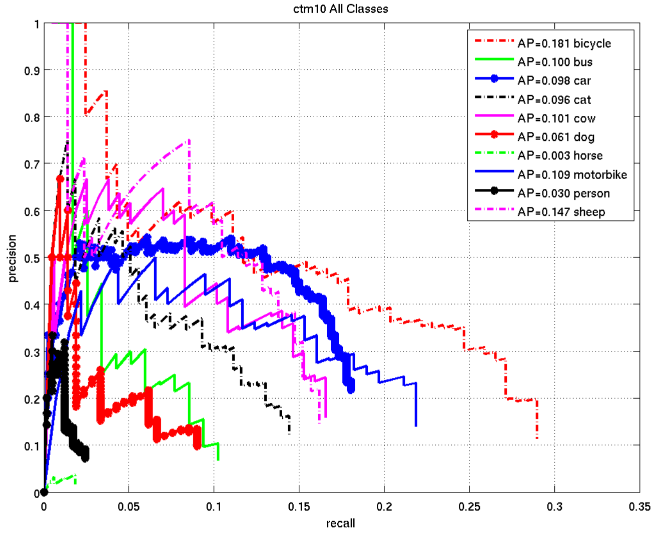

Figure 12: CTM precision-recall curves and Average Precision.

based techniques) utilize classifiers which are capable of predicting a topic distribution from just one segment. In the CRF literature, context modeling is achieved by coupling the unary classifiers via pairwise potentials.

A severe limitation of our approach was the lack of spatial constraints used in the bag-of-words model. Our model discarded information about neighboring segments inside one segmentation and overlapping segments across different segmentations. Consider a hypothetical situation where we have an image of a person with blue pants and a red shirt standing on some green grass with some blue sky above the person. In the case that the shirt and pants were never grouped into one segment, the blue pants would be probably associated with the sky class even if they were surrounded by green and red segments. This is an example of the lack of context.

6.2 Multiple Segmentations Discussion

As opposed to the work of Russell *et al.* [11] – where the goal was to find good segments – our work used all of the segments in tandem. This means that in our work segments sometimes represented objects as a whole and sometimes an object of interest was represented as a collection of segments.

7 Future Work

- *Detection Confidence Measure*: The mean topic response didn't work so well as a confidence (can be seen by the jaggedness of the precision recall curve). Two ideas to improve this measure are 1.) incorporating some type of spatial voting where each segment votes for the object center and the density of the votes is a confidence and 2.) after obtaining the region for a particular class we could match the histogram of words falling in that region to the histograms of words associated with all instances in the database (the distance to the nearest neighbor of that class would be inversely proportional to the confidence).

- *Intelligent Lexicon Creation*: As opposed to naively running k-means in an unsupervised fashion, class labels could have been used to choose words that are more discriminative with respect to class labels.
- *CTM versus LDA*: The Correlated Topic Model has been shown to outperform Latent Dirichlet Allocation for unsupervised topic discovery in text corpora. We would like to answer why the Correlated Topic Model did not outperform Latent Dirichlet Allocation for our particular application.
- *Segmentations*: How much does the image-driven bottom-up segmentation help up? Perhaps simply using many partitions of the image into box regions which doesn't rely on the image data could be sufficient for the localization of objects within a bounding box.

8 Conclusion

In conclusion, we incorporated multiple segmentations into Latent Topic Models and showed that we can get localization and segmentation of objects. Due to the availabilty of labeled data, we were able to train our models in a supervised fashion as opposed to the more commonly used unsupervised training procedure. Finally, we compared Latent Dirichlet Allocation and the Correlated Topic Model and found that they gave similar results, with LDA slightly outperforming CTM.

References

[1] J Aitchison and S.M. Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67, 1980.

[2] David Blei and John Lafferty. Correlated topic models. *Advances in Neural Information Processing Systems*, 18, 2005.

[3] David Blei, Andrew Ng, and Michael Jordan. Latent dirichlet allocation. *Advances in Neural Information Processing Systems*, 14, 2002.

[4] G. Csurka, C. Bray, C. Dance, and Lixin Fan. Visual categorization with bages of keypoints. 2004.

[5] Li Fei-Fei and Pietro Perona. A hierarchical bayesian model for learning natural scene, categories. In *IEEE Computer Science Society International Conference of Computer Vision and Pattern Recognition*, October 2005.

[6] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning, 2003.

[7] T. Griffiths and M. Steyvers. Finding scientific topics. In *Proceedings of the National Academy of Sciences*, 2004.

[8] Derek Hoiem, Alexei A. Efros, and Martial Hebert. Geometric context from a single image. In *International Conference of Computer Vision (ICCV)*. IEEE, October 2005.

[9] Michael I. Jordan, Zoubin Ghahramani, Tommi Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

[10] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. 18th International Conf. on Machine Learning*, pages 282–289. Morgan Kaufmann, San Francisco, CA, 2001.

[11] Bryan C. Russell, Alexei A. Efros, Josef Sivic, William T. Freeman, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. In *Proceedings of CVPR*, June 2006.

[12] Josef Sivic, Bryan Russell, Alexei Efros, Andrew Zisserman, and William Freeman. Discovering object categories in image collections. Technical Report AIM-2005-005, Massachusetts Institute of Technology, 2005.

[13] Michel Vidal-Naquet and Shimon Ullman. Object recognition with informative features and linear classification. In *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 281, Washington, DC, USA, 2003. IEEE Computer Society.

[14] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision*, pages 1800–1807, Washington, DC, USA, 2005. IEEE Computer Society.

[15] Stella Yu. *Computational Models of Perceptual Organization*. PhD thesis, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, May 2003. CNBC and HumanID.