Capturing Spatial Context in Images with a Relational Dictionary

Roman Stanchak
Department of Computer Science
University of Maryland
College Park, MD 20742
roman@cs.umd.edu

Abstract

Probabilistic generative models such as Probabilistic Latent Semantic Indexing (PLSI), Latent Dirichlet Allocation (LDA), and their variants have been used with some success on the task of object recognition in images. Recent approaches integrate spatial information into the generative process leading to complex hierarchical models. This work takes a simpler approach by encoding the neighborhood of a particular visual word in a relation dictionary. Thus, an image is represented as histogram over this relational dictionary, and standard topic modeling approaches can be applied. We examine three methods of inducing the neighborhood graph and compare them on an image recognition task using the LabelMe benchmark database. Results for the relational dictionary are on-par with a standard non-relational dictionary, but consistently poor, thereby suggesting a flaw in the experimental setup.

1 Introduction

The task of automated image understanding involves identifying the objects of which an image consists, how they might be interacting, and the general context of the image. Each of these components are inter-related, for instance, a city street scene may consist of cars, a road, buildings and people. These objects all have typical relationships: cars are on the road, people are on the sidewalk, and buildings adjoin with sidewalks and are behind people. Recently, generative probabilistic models, namely *Topic Models* such as *Probabilistic Latent Semantic Indexing*[8] and *Latent Dirichlet Allocation*[1] have been used with some success on the task of image understanding[4][2].

Traditional topic modeling methods generally assume a 'bag of features' model in which there is no notion of order, location or distance associated with each *feature* in the model. In the real world, order plays a large role in defining the context in which features can appear. For instance, the meaning of a textual document is lost if the words are randomly permuted. Similarly, an image of a human face ceases to be recognizable if the relative positions of the eyes, nose and mouth are scrambled. Although topic models have seen much success in their application to various domains, these small examples suggest that there is something to be gained by considering the spatial relationships between features.

2 Background

A core task of general image understanding is object recognition, that is, determining which of some number of known objects an image consists of. The typical approach to object recognition involves several standard steps[3]:

1. Detection of interest points x_i in the image.
2. Description of local image region around x_i.
3. Quantization of the descriptor into a fixed size dictionary $\mathcal{D}$ of *visual words*.
4. Application of a classifier.

Prior to application of a classifier, recent approaches to object recognition have found success using probabilistic generative models to cluster the distribution of *visual words* associated with a particular object class label. For instance, [4] uses a variant of the Author-Topic Model[13] to model natural scenes. This work uses the bag of words model that ignores the spatial location associated with each *visual word*. Probabilistic generative models have been used to model feature co-occurrence [21], the relative locations of features [17] [22] and the spatial transformations that led a particular configuration [16]. Other approaches to object recognition that utilize spatial relationships are: [5] uses EM to learn a Bayesian model of appearance, shape and scale; [10] aggregates statistics of local appearance at multiple scales; and both [15] and [2] integrate hierarchical image segmentation within a topic modeling framework. Spatial relationships have also been considered in the general context of topic modeling. Wallach[20] describes a bigram model for text corpora, Griffiths and Steyvers [7] describe LDA Collocation which also considers bigrams, and LDA-Composite, which overlays a Hidden Markov Model (HMM) on the observed sequence of words. These works consider only the sequential order of words, and do not explicitly extend to multi-dimensional spaces. Finally, Wu[24] discusses the integration of spatial constraints into a Dirichlet process using Markov Random Fields.

3 Approach

In this work, we consider encoding the spatial relationship among neighboring codewords in a secondary *relational dictionary* $\mathcal{R}$, and then using $\mathcal{R}$ as input to a topic modelling algorithm.

3.1 Relational Dictionary

The relational dictionary consists of all unordered pairs of codewords in the standard dictionary (see Figure 1). Thus, a standard codeword dictionary of size d produces a relational dictionary of size $d \cdot (d+1)/2$. An image is represented as histogram over the relational dictionary $\mathcal{R}$ by projecting all edges in the neighborhood graph onto $\mathcal{R}$, and simply counting how many times each relational codeword r occurs. In summary, the process for producing a relational representation of is as follows:

1. Detect interest points, compute descriptors, and quantize into *visual words*.
2. Induce a graph on the detected interest points x_i
3. Project neighbors onto relational dictionary

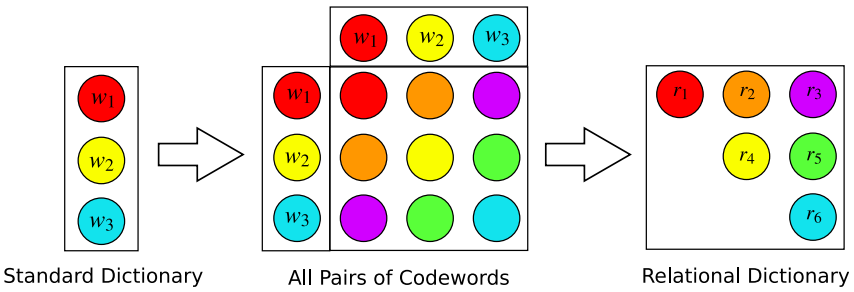

Figure 1: Relational Dictionary Formation. From right to left: Descriptors computed on detected interest points are quantized into a fixed size dictionary (of length 3 here). The relational dictionary is formed by all unordered pairs of standard codewords.

3.2 Inducing a Graph

In order to determine the distribution of relational codewords for each image, some notion of neighborhood between feature positions x_i must be defined. Intuitively, we would like the distribution to reflect (a) that interest points from the same object should be related (b) interest points from an object and the background it typically occurs in should be related. In both cases, the simplest general method which satisfies both properties is that neighbor interest points should have nearby spatial positions. Here we compare three possibilities for inducing a graph among the positions x_i:

- K-Nearest Neighbors
- Fixed-Radius Neighbors
- Delaunay Triangulation

For K-Nearest Neighbors, the neighborhood for the point x_i consists of the k instances of x_j with the minimum L2 distance $||x_i - x_j||$. For Fixed-Radius, $e_{ij} = 1$ if $||x_i - x_j|| < r$, for some fixed r. Finally, the Delaunay Triangulation produces a triangulation of the points $\{x_i\}$, (so the resulting graph is fully connected graph) based on the principle of maximizing the minimum angle of all the angles of the triangles in the triangulation [23]. The rationale for choosing these methods was simplicity and readily available implementations. The graphs induced by each method are shown for an example image in Figure 2.

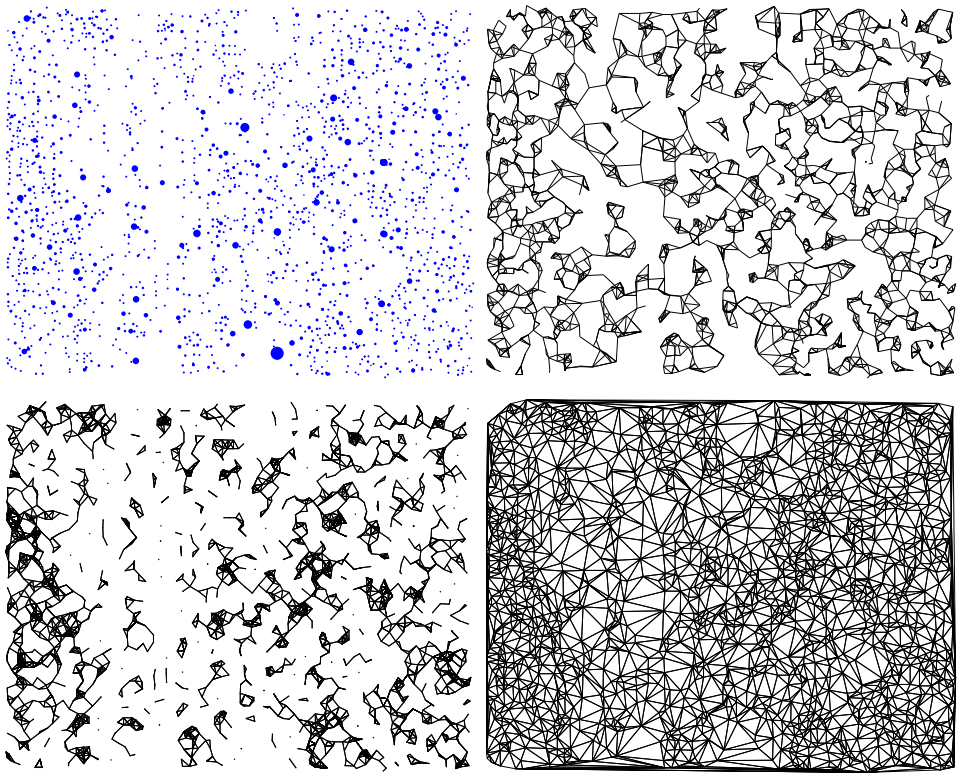

Figure 2: From left to right, top: Location of detected features, K-Nearest Neighbors(k=5), bottom: Fixed Radius Neighbors(r=10 pixels), Delaunay Triangulation.

4 Evaluation

The graph induction methods were evaluated in terms of a binary image classification task. That is, for each image, the goal is to predict whether a given tag occurs in the image. Two methods for doing this were evaluated. In the first, Latent Dirichlet Allocation[1] is applied to the merged

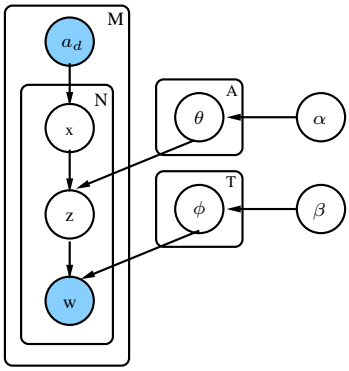

Figure 3: The Author Topic Model. In the Object-Image interpretation, the generative process is as follows: Each of M images has an associated set of observed object tags $\mathbf{a}_d$. Each of the A possible object tags is associated with a distribution θ of topics. Similarly, each of the T possible topics is associated with a distribution ϕ over the visual words in the image. For each word w, a tag x is sampled uniformly from $\mathbf{a}_d$, then a topic z from θ_x, then the word w from ϕ_z. α and β are hyper-parameters of the Dirichlet distribution from which ϕ and θ are drawn.

training and test set to learn a set of topic mixture components for each image. The class labels are not used in this process. The topic mixture components can be viewed as a reduced dimensionality representation of the original data. This representation is then used as the input data to a standard Machine Learning classification algorithm such as Naive Bayes or Support Vector Machines.

In the second method, the Author-Topic Model[13] variant of LDA is used. Here, the class labels are used in the generative model to influence the distribution of topics for each image (See Figure 3. Hence, only the training set is used to learn the topic mixture components. Rather than inferring the mixture components for the test set and using a classifier, the generative model can be used directly to infer the class labels most likely to generate the observed test data. This value was calculated as follows for the test data based on the estimated object-topic and topic-word assignments in the training data:

$$P(\mathcal{O}|\mathcal{I}) = \prod_{w \in \mathcal{I}} \sum_{z} P(z|\mathcal{O}) P(w|z)$$

For both LDA and Author Topic, Gibbs Sampling was used to infer the mixture parameters based on code provided in [6].

4.1 Data

LabelMe is a web-application that allows casual users to upload photos and then annotate them with tagged polygons [14]. At the time of writing, the full database consists of 162993 images, 43182 of which have at least one annotation. A subset of the database is available for benchmarking purposes. The training set contains 2920 images, 32164 total annotations and 127 unique tags. The test set contains 1033 images and 32853 total annotations. The images are primarily street scenes, so they tend to contain a large proportion of tags such as *sidewalk*, *car*, *building*, *road*, etc.

4.2 Representation

Each image was represented in a standard dictionary of 100 and 10000 codewords, and relational dictionaries built using the standard dictionary of size 100 (resulting in a relational dictionary size of 5500). The Harris-Affine interest point detector and SIFT[11] descriptor provided by [19] were used to find and describe the interest points. The dictionary was learned from the training data using code provided by [18], which is based on a vocabulary tree described in[12].

5 Results

Receiver-Operator-Characteristic (ROC) curves comparing the graph induction techniques are shown in Figure 4 for two image categories. *Unfortunately, at the time of this writing, the experiments utilizing the Author Topic model described previously failed to complete*. As can be seen the plots, classification accuracy is disappointingly low. For all methods used, the True/False Positive ratio was only slightly better than random chance, and in some cases, worse. Interestingly, the Delaunay method produced drastically different results than the others, which all had similar performance. While in some cases it appeared to improve performance, in others it degraded.

The baseline LDA approach has been demonstrated to be effective on other image databases [4], so it is suspicious that the baseline does poorly for the LabelMe benchmark database. Other approaches using the LabelMe benchmark database report significantly results, but use more sophisticated generative models than the baseline LDA. Further experimentation is needed to either (a) confirm that baseline LDA is ill-suited to the LabelMe database or (b) find an error in the experimental procedure that led to these poor results.

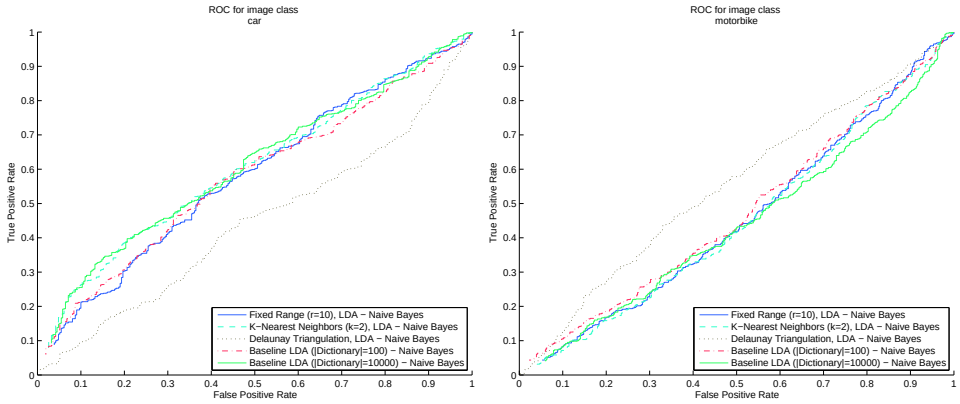

Figure 4: ROC curves comparing the graph induction techniques with baseline for two image categories (car and motorbike).

6 Conclusion and Future Work

This work examined the use of relational codewords for encoding spatial relationships in an image for the task of object recognition. We compared three methods of inducing a graph on a sparse set of detected features in an image: K-Nearest Neighbors, Fixed-Radius Neighbors, and Delaunay Triangulation. This relational dictionary was compared to a standard dictionary on an binary recognition task. Results were inconclusive. The relational dictionaries performed similarly to the standard dictionary, but performance was poor across the boards. Further experimentation is needed to confirm the results. One possible direction for future research is examining more sophisticated methods for graph induction. For instance, one intriguing alternative is integrating a probabilistic generative edge model into the topic modelling framework. This would avoid the quadratic explosion in the number of relational codewords necessitated by comparing all pairs of standard codewords, and practically allow a greater number of standard codewords to be used. Finally, there is some evidence that suggests dense sampling of visual words is superior to the sparse interest point detection used here[9]. Dense sampling would lead to a straightforward definition of neighborhood as a standard 2D uniform grid. Such a formulation may allow the adaptation of the technique described in [24], which uses Markov Random Fields.

References

[1] D. Blei, A. Ng, and M. Jordan. Latent dirichlet allocation. *Journal of machine Learning Research 3*, 2003.

[2] Liangliang Cao and Li Fei-Fei. Spatially coherent latent topic model for concurrent segmentation and classification of objects and scenes. *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, 14-21 Oct. 2007.

[3] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. *Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[4] L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. *Proc. CVPR*, 5, 2005.

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 2:II–264–II–271 vol.2, 18-20 June 2003.

[6] T.L. Griffiths and M. Steyvers. Matlab topic modeling toolbox. Online; `http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm`; accessed 05/15/2008.

[7] TL Griffiths, M Steyvers, DM Blei, and JB Tenenbaum. Integrating topics and syntax. In *Advances in Neural Information Processing Systems 17*, 2005.

[8] Thomas Hofmann. Probabilistic latent semantic analysis. In *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[9] F. Jurie and B. Triggs. Creating Efficient Codebooks for Visual Recognition. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 1, 2005.

[10] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *cvpr*, 02:2169–2178, 2006.

[11] D.G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[12] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. *Proc. CVPR*, pages 2161–2168, 2006.

[13] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 487–494, 2004.

[14] B.C. Russell, A. Torralba, K.P. Murphy, and W.T. Freeman. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, 77(1):157–173, 2008.

[15] Bryan C. Russell, William T. Freeman, Alexei A. Efros, Josef Sivic, and Andrew Zisserman. Using multiple segmentations to discover objects and their extent in image collections. *cvpr*, 02:1605–1614, 2006.

[16] E. Sudderth, A. Torralba, W.T. Freeman, and A. Willsky. Describing visual scenes using transformed dirichlet processes. *Advances in Neural Information Processing Systems*, 19, 2006.

[17] Erik B. Sudderth, Antonio Torralba, William T. Freeman, and Alan S. Willsky. Learning hierarchical models of scenes, objects, and parts. *iccv*, 2:1331–1338, 2005.

[18] Andrea Vedaldi. Bag of features. Online; `http://vision.ucla.edu/~vedaldi/code/bag/bag.html`; accessed 05/14/2008.

$\mathcal{I}$	An image
$\mathcal{O}$	An object class
x_i	Location of interest point i in image coordinates
d_i	Descriptor of interest point i
w_i	Codeword corresponding to Descriptor d_i.
$\mathcal{N}_i$	Neighbors of interest point i.
e_{ij}	Edge exists between i and j. $e_{ij} = 1$ if $j \in \mathcal{N}_i$, 0 otherwise
$\mathcal{D}$	Dictionary of codewords
$\mathcal{R}$	Relational dictionary
r_{ij}	Relational codeword corresponding to the edge i,j
z	Latent topic variable

Figure 5: List of Notation

[19] Andrea Vedaldi. Sift++. Online; http://vision.ucla.edu/~vedaldi/code/siftpp/siftpp.html; accessed 05/14/2008.

[20] Hanna M. Wallach. Topic modeling: beyond bag-of-words. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 977–984, New York, NY, USA, 2006. ACM.

[21] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. *IEEE CVPR*, 2006:1597–1604, 2006.

[22] Xiaogang Wang and Eric Grimson. Spatial latent dirichlet allocation. In *Proceedings of Neural Information Processing Systems Conference*, 2007.

[23] Wikipedia. Delaunay triangulation — Wikipedia, the free encyclopedia, 2008. Online; http://en.wikipedia.org/wiki/Delaunay_triangulation; accessed 05/14/2008.

[24] Liang Wu, Predrag neskovic, and Luiz Pessoa. Dirichlet process mixture model with spatial constraints. Technical Report IBNS Technical Report 2007-02, Institute for Brain and Neural Systems - Brown University, Providence, RI, 2007.