Kernel Density Topic Models: Visual Topics Without Visual Words

Konstantinos Rematas
K.U. Leuven
ESAT-iMinds
krematas@esat.kuleuven.be

Mario Fritz
Max Planck Institute for Informatics
mfrtiz@mpi-inf.mpg.de

Tinne Tuytelaars
K.U. Leuven
ESAT-iMinds
tinne.tuytelaars@esat.kuleuven.be

Abstract

The computer vision community has greatly benefited from transferring techniques originally developed in the document processing domain to the visual domain by means of discretizing the features space into visual words. This paper reinvestigates the necessity of this artificially discretization of the continuous space of visual features and consequently proposes an alternative formulation of the popular topic models that is based on kernel density estimates. Results indicate the benefits of our model in terms of decreased perplexity as well as improved performance on object discovery tasks.

1 Introduction

Computer vision has been greatly inspired by techniques originally developed for text analysis and document processing. Most prominently the bag-of-words representations and its derivates are still to date one of the most successful techniques for visual categorization. Another example are topic models that have boosted unsupervised learning of visual representations [5, 7] and led to new methods for discovering object classes [12, 6] in a data-driven manner.

All these methods are based on an analogy between the visual and the text domain. However, while text naturally decomposes into words, visual features need to be clustered (discretized) into visual words in order to complete the analogy. Different methods from hard (*e.g.* K-means) to soft (*e.g.* Gaussian mixtures) quantization have been studied. This discretization step can be criticized in many ways. It is not only counter-intuitive to depart from the continuous nature of the popular visual descriptors, but it also causes a loss of valuable information [3]. On top of that, we observe cases where the result is critically dependent on the choice of discretization parameters as well as cases where we simply do not find a satisfactory discretization.

This paper reevaluates the necessity of imposing a somewhat artificial discretization of the continuous space of visual features for the widely used topic models for visual learning and explore ways how to define them directly on continuous feature spaces. To this end, we replace the discrete multinomial distribution by a kernel density estimate (KDE). Our results show that in this manner we can still benefit from the desirable properties of grouped clustering for which topic models attained fame in the computer vision community without postulating any strong assumptions on the underlying density and – most importantly – avoiding the discretization of the visual feature space.

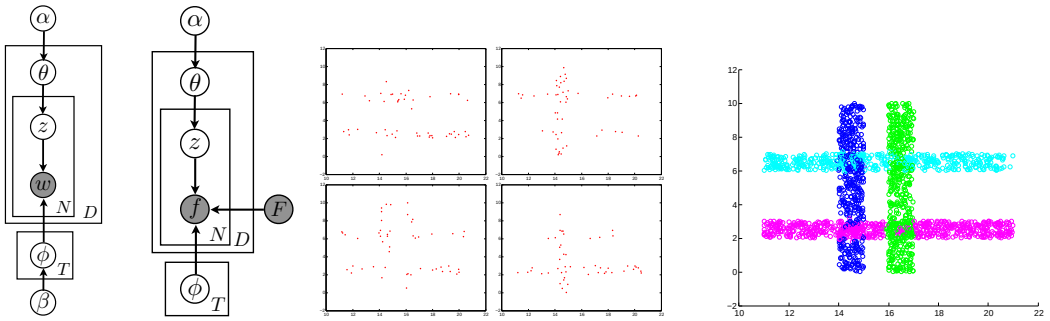

Figure 1: Left: graphical models of LDA and our wordless model. Center: 4 documents with samples from different topics. Right: approximation obtained by our model of the topic distributions that generated the samples.

Related Work. Probably most related to our work are [1] and [10]. Both propose a probabilistic mixture model for visual words on a continuous features space. In [1], the authors perform segmentation in one image or in a collection of images based on the affinity between pixel features. The main difference is that their approach does not consider the co-occurrence of features across documents as in a topic model but only their similarity in feature space. In [10], since the objective is the vocabulary generation, the topics are still distributions over words but each word is considered itself as a Gaussian distribution over features and the vocabulary generation is performed together with model inference.

2 Traditional Visual Words and Topic Models

Visual Vocabulary Generation. The first step of the vocabulary generation pipeline is to extract sets of local features (patches) from the images. Then vector quantization clusters image features into a predefined number of clusters. The centers of the clusters correspond to visual words and all features are assigned to the word that is closest in feature space. An image is represented as a collection of visual words.

This process may seem straightforward and easy to implement and apply. However several issues arise: i) What is the correct vocabulary size? ii) Dependence on random initialization of the clustering algorithm and iii) Loss of discriminative power due to vector quantization as shown in [3].

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) [2] is a generative model originally proposed for collections of text documents. A document d is described as a distribution θ over T topics, with a topic being a distribution ϕ over V words from a fixed vocabulary. The joint distribution of the LDA model (see Figure 1) is:

$$P(\mathbf{w}, \mathbf{z}, \theta, \phi | \alpha, \beta) = P(\phi|\beta)P(\theta|\alpha)P(\mathbf{z}|\theta)P(\mathbf{w}|\phi_z), \tag{1}$$

where $\mathbf{w}$ are the observed words, $\mathbf{z}$ their topic assignments and α and β the Dirichlet priors on θ and ϕ respectively. Given a corpus D of documents, the goal is to estimate for every document d the distribution over topics θ_d and for every topic j its distribution over words ϕ_j. For inference we use Gibbs sampling as in [8].

3 A Wordless Topic Model

The generative process of our model is similar to the one of [2, 8] described above, except that the multinomial parameterization of the discrete visual word distribution is replaced by a kernel density directly defined on the visual feature space:

- For every document $d \in D$,
  - sample a topic distribution θ^d from a Dirichlet with hyper-parameter α.
- For every image feature $f_i \in d$,
  - sample a topic z_i from $Mult(\theta^d)$.

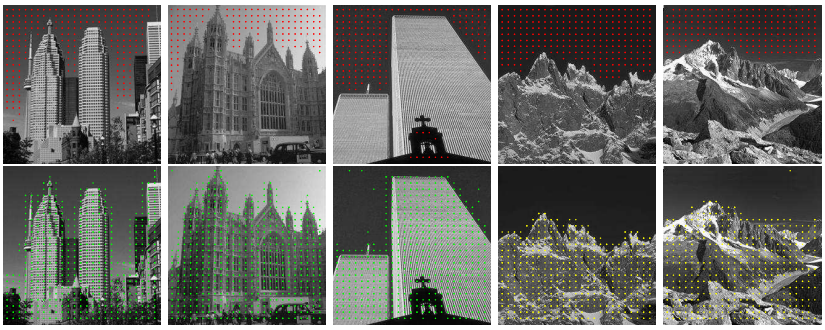

Figure 2: Visualization of 2 out of the 6 topics for the Tall Building and Mountain classes from the 15 Scenes dataset [5]. Such decomposition of features allows class specific topics (*e.g.*building vs mountain) and sharing topics (*e.g.*sky).

– sample the feature from the topic-specific kernel density estimate given by the subset of the training data F specified by the binary indicators ϕ_{z_i}.

Figure 1 (left) illustrates the graphical model of LDA and our wordless model. The probability of feature f_i coming from topic $z_i = j$ can be found by applying kernel density estimation (Gaussian kernel) on the set of features that populate the topic distribution ϕ_j.

Inference through Gibbs sampling. Instead of trying to estimate θ^d and ϕ_j directly, we estimate the topic assignment z_i for every feature f_i using Gibbs sampling [8]. In short, firstly the topic assignments of the features are randomly initialized. Then, for each feature a topic is sampled from the conditional distribution $P(z_i = j|\mathbf{z}_{-i}, \mathbf{f})$, where $\mathbf{z}_{-i}$ are all topic assignments except for feature f_i and $\mathbf{f}$ all the features in the dataset. This process is performed for every feature for a certain number of iterations. The aforementioned conditional can be decomposed as:

$$P(z_i = j|\mathbf{z}_{-i}, \mathbf{f}) \propto P(f_i|z_i = j, \mathbf{z}_{-i}, \mathbf{f}_{-i})P(z_i = j|\mathbf{z}_{-i}) \tag{2}$$

The first term is the probability of a feature f_i being assigned to topic j given all other topic assignments and features, while the second term is the probability of a certain topic given all other topic assignments. With our KDE model this becomes:

$$P(f_i|z_i = j, \mathbf{z}_{-i}, \mathbf{f}_{-i}) \propto \frac{1}{|\Phi_{j,-i}|} \sum_{r}^{n} \Phi_{j,-i} \cdot K(f_i, f_r). \tag{3}$$

In the above, $\Phi_{\cdot,\cdot}$ is a $n \times T$ binary matrix where n is the total number of features and T the number of topics. If the $\Phi(i, j)$ is one, this indicates that feature f_i was assigned to topic $z_i = j$. In the above equation $-i$ indicates that we do not consider the leave-out feature.

The second term of Eq. 2 can be estimated from the number of features that were assigned to a specific topic and appear in a certain document. From [8] we have: $P(z_i = j|\mathbf{z}_{-i}) = n^{d_i}_{-i,j} + \alpha$, where $n^{d_i}_{-i,j}$ is the total number of features in document d_i that were assigned to topic j, not including feature f_i and α is the Dirichlet hyper-parameter. In other words, this term represents the probability of assigning topic $z_i = j$ to a particular document. As in [8], $\theta^d_j = n^d_j + \alpha$, but for the estimation of ϕ_j we use the kernel density estimate from Eq. 3 with all features observed.

We illustrate the capability of this model to decompose feature sets by running an initial experiment on a subset of scene 15 dataset. Figure 2 shows how the model discovers different regions such as sky, building, mountain in an unsupervised manner and without vector quantization.

4 Experiments

Unsupervised Topic Discovery. In order to quantitatively measure the ability of our model to discover meaningful topics we apply the model in a collection of images that come from 4 different classes (Faces, Motorbikes, Airplanes and Cars) from Caltech 101 [4]. We choose 100 images for every class and we assign each image d to topic j according to $arg\max_j \hat{\theta}^d_j$. We set the number of

k \ σ	0.6	0.7	0.8	0.9
20	94.2%	94.2%	94.5%	94.0%
25	95.0%	95.0%	95.0%	94.7%
30	93.5%	97.2%	97.5%	97.2%

Voc size	%
1000	93.3 ± 5.5
2000	92.9 ± 2.7
3000	80.9 ± 11

Table 1: Accuracy of our model for different sigma and k nearest neighbors settings (left) against standard LDA with different vocabulary size and different runs (right) using SIFT.

k \ σ	1.8	2	2.2	2.4
20	94.0%	94.0%	94.0%	94.0%
25	96.5%	98.2%	97.5%	97.5%
30	96.5%	99.0%	96.5%	96.2%

Voc size	%
500	84.1 ± 8.7
750	87.8 ± 9.0
1000	85.2 ± 10.7

Table 2: Accuracy of our model for different sigma and k nearest neighbors settings (left) against standard LDA with different vocabulary size and different runs (right) using NIMBLE.

topics T to 4 and the hyper-parameter α to $50/T$. We use two different types of features, SIFT [11] and NIMBLE [9]. Moreover, we compare our method with standard LDA using a fixed vocabulary. We test with different vocabulary sizes and for every size we repeat the clustering 5 times.

We have two free parameters from the kernel density estimator: the bandwidth σ and the number k of nearest neighbors we use to efficiently approximate the KDE. The intervals for σ can be found simply by observing the distance distribution between the features without any other knowledge.

In Table 1 (left) we show the accuracy of our model by keeping one parameter constant and varying the other for SIFT features. As we can see the results for different σ and k do not vary considerably and they are better than the standard LDA results (Table 1 right), which present also high variance.

Table 2 (left) shows the performance of our model using NIMBLE features. Since the NIMBLE features are stronger $w.r.t.$ class-information capture, the performance is higher than SIFT. Also here the variance of the accuracy is low for our model for different σ and k, similar with the SIFT experiment and, compared with the standard LDA, our model presents higher performance.

Perplexity. A standard measure to evaluate the generalization ability of a model is the perplexity $perpl(D_{test}) = \exp -\frac{\sum_{d=1}^{M} \log P(\mathbf{f}_d|d)}{\sum_{d=1}^{M} N_d}$, with M the number of testing documents, N_d the number of features in the d test document, $\hat{\phi}$ the topic specific distributions estimated on a training set and $\tilde{\theta}^d$ the topic distribution for the testing document d. For computational reasons we approximate the likelihood of the KDE by retrieving 0.1% of the nearest neighbors.

In Figure 3 we present the perplexity of our model compared with standard LDA for different number of topics in the previous dataset. Our model has lower perplexity than standard LDA for different vocabulary sizes. In addition, the perplexity is even less dependent on the choice of the number of topics and remains low for a broad range of σ.

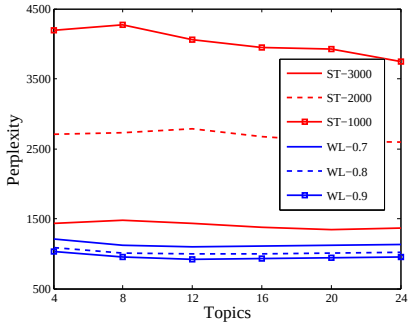

Figure 3: Perplexity of our model for different sigma (blue) and standard LDA for different vocabulary sizes (red).

5 Conclusion

In this paper we introduced an alternative to the popular visual word representation in the context of topics models. By replacing the multinomial mixture model of the Latent Dirichlet Allocation with a non-parametric kernel density estimate, we have proposed a novel wordless topic model. On synthetic as well as real data we show that the model maintains the desired grouped clustering properties and achieves a decomposition of images into visual topics without the need of visual words and the associated discretization.

References

[1] M. Andreetto, L. Zelnik-Manor, and P. Perona. Non-parametric probabilistic image segmentation. *ICCV*, 2007.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 2003.

[3] O. Boiman, E. Shechtman, and M. Irani. In defense of nearest-neighbor based image classification. *CVPR*, 2008.

[4] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1), 2007.

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

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

[7] M. Fritz and B. Schiele. Decomposition, discovery and detection of visual categories using topic models. *CVPR*, 2008.

[8] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Science*, 2004.

[9] C. Kanan and G. Cottrell. Robust classification of objects , faces , and flowers using natural image statistics. *CVPR*, 2010.

[10] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization and segmentation. *Image and Vision Computing*, 2009.

[11] D. G. Lowe. Object recognition from local scale-invariant features. *ICCV*, 1999.

[12] J. Sivic, B. Russell, A. A. Efros, A. Zisserman, and B. Freeman. Discovering Objects and Their Location in Images. *ICCV*, 2005.