Universität Augsburg
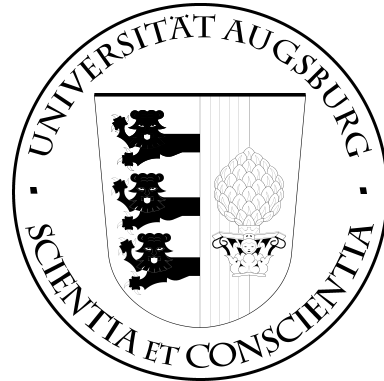
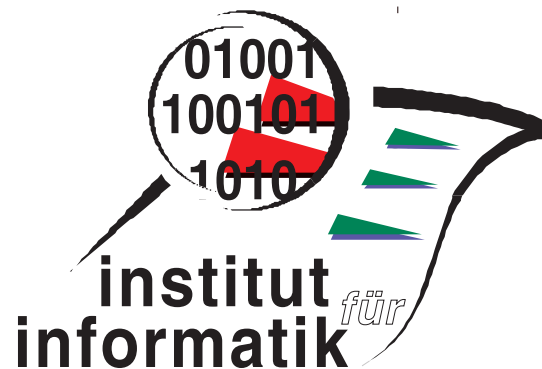# Correlated Topic Models for Image Retrieval

T. Greif, E. Hörster, R. Lienhart

Institut für Informatik
D-86135 Augsburg

Correlated Topic Models for Image Retrieval

Thomas Greif
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
thomas.greif@student.uni-augsburg.de

Eva Hörster
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
hoerster@informatik.uni-augsburg.de

Rainer Lienhart
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
lienhart@informatik.uni-augsburg.de

ABSTRACT

In our previous work [4] we have shown that the representation of images by the Latent Dirichlet Allocation (LDA) model combined with an appropriate similarity measure is suitable for performing large-scale image retrieval in a real-world database. The LDA model, however, relies on the assumption that all topics are independent of each other – something that is obviously not true in most cases. In this work we study a recently proposed model, the Correlated Topic Model (CTM) [1], in the context of large-scale image retrieval. This approach is able to explicitly model such correlations of topics. We experimentally evaluate the proposed retrieval approach on a real-world large-scale database consisting of more than 246,000 images and compare the performance to related approaches.

1. INTRODUCTION

Very large image repositories such as Flickr become more and more popular, because they offer an easy way to share images among people. As the number of uploaded pictures grows, techniques to organize and categorize these images are needed more than ever. Indexing is commonly realized based on image tags provided by the authors of the images or the community. Usually no constraints are imposed on what could be a tag making them largely subjective. Sometimes they do not even reflect the visible image content.

In previous works [5][4] we have exploited directly the image content in order to find similar images as perceived by humans. Both works use topic models, probabilistic Latent Semantic Analysis (pLSA [3]) in [5] and Latent Dirichlet Allocation (LDA [2]) in [4], to derive a higher level image representation appropriate for retrieval. Topic models originate from modeling large databases of text documents. When applied to images instead of documents, each topic can be thought of as a certain object type that is contained in an image. The topic distribution then refers to the degree to which a certain object/scene type is contained in the image. In the ideal case, this gives rise to a low-dimensional description of the coarse image content and thus enables retrieval in very large databases. Another advantage of such models is that topics are learned automatically without requiring any labeled training data.

However, these models rely on the wrong assumption that all topics are independent of each other. For example, a text document about probability theory is most likely also about statistics and not about architecture. Or in the visual domain, one would assume that if one sees a car that probably a street is depicted, too.

In this work we explore the recently proposed CTM [1], a model developed for text document analysis and closely related to the LDA. In contrast to the LDA it does not rely on the independence assumption of the topics. We describe how this model can be applied to the image domain and we evaluate whether the CTM is appropriate for modeling large-scale image databases for image retrieval. The approach is evaluated experimentally on a real world database consisting of more than 246,000 images.

2. CTM-BASED IMAGE RETRIEVAL

2.1 The Correlated Topic Model

As the CTM has been originally developed in the text domain, we will first review its generative model for the case of a text corpus. In Section 2.2 we will then describe how we apply the model to images.

Much like in the LDA model, CTM [1] assumes that each document is composed of words that all arise from mixtures of *topics*, i.e. documents are represented by finite mixtures over hidden topics and in turn each topic is characterized by a distribution over the entire vocabulary. Unlike the LDA, where the topic proportions of a specific document are drawn from a Dirichlet and where therefore the correlation between different topics is disregarded, the CTM draws these topic proportions from a logistic Normal distribution. That means in detail, to generate the topic proportions for a document, a random vector is drawn from a multivariate Gaussian and then mapped to the simplex to obtain a multinomial parameter. Thus, the covariance of the Gaussian entails dependencies between the elements of the vector.

Assuming each document $\mathbf{w}_d$ of a corpus/database of D documents is composed of a sequence of N_d words, i.e. $\mathbf{w}_d$, is a vector of dimension N_d containing the words w_i a document d consists of. Each word in a document is associated
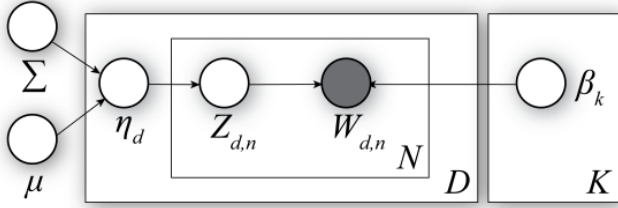
Figure 1: The graphical representation of the CTM. D denotes the number of documents, K the number of hidden topics and N the number of words each document is composed of. The gray node shows the only observable random variable $w_{d,n}$.

with one of the K topics in the model. According to [1], the generative process an N-word document d arises from, can formally be summarized as follows [1]:

1. Draw $\eta_d | \{\mu, \Sigma\} \sim N(\mu, \Sigma)$, where μ denotes a K-dimensional mean vector and Σ a covariance matrix of size $K \times K$

2. For $n \in \{1, \ldots, N_d\}$:

    (a) Draw topic assignment $z_{d,n} | \eta_d$ from $\mathrm{Mult}(f(\eta_d))$

    (b) Draw word $w_{d,n} | \{z_{d,n}, \beta_{1:K}\}$ from $\mathrm{Mult}(\beta_{z_{d,n}})$

and where $f(\eta)$ denotes a mapping of the natural parameterization of the topic proportions to the mean parameterization.

$$\theta = f(\eta) = \frac{\exp\{\eta\}}{\sum_i \exp\{\eta_i\}} \qquad (1)$$

The graphical representation of the CTM is shown in figure 1.

The only observable variables in the CTM are the words each document consists of. Learning the parameters of such a model given a set of training documents is accomplished by a variational expectation-maximization (EM) procedure. Given the learned model we can estimate the topic proportions of a new document by a variational inference algorithm. Details regarding the learning and inference algorithms in the CTM model can be found in [1].

2.2 Image Representation

Since we are interested in modeling image databases instead of text databases, the documents in our model correspond to images. Hence, the topic proportions of a document correspond to a mixture of different objects/scene types in an image.

In order to apply the above described CTM model to images we need to represent each image by a set of *visual words*. We define visual words as vector-quantized local image descriptors and the finite set of visual words constitutes the discrete visual vocabulary.

Thus the first step in our approach is to generate a visual vocabulary. We learn such a vocabulary by extracting local image descriptors from the images in the database, each represented by a feature vector, and then applying k-means clustering. The cluster centers are chosen as visual words. K-means clustering is computationally very expensive when using a large number of (high-dimensional) feature vectors and a large number of clusters. Thus we instead compute smaller vocabularies on non-overlapping subsets of the entire feature set and subsequently merge the resulting words into one final vocabulary.

As local image descriptors we use the well-known SIFT features [6] extracted at extrema of the difference of Gaussian pyramid. It should be noted that each image may compute a different number of features depending on the structure and texture of the image content. Each extracted local feature has 128 dimensions.

Having computed the vocabulary, the images are represented by *word count vectors* by replacing each detected feature vector with its most similar visual word in the vocabulary. The most similar word is defined as the visual word which is closest in the 128-dimensional vector space. Counting the number of occurrences of each visual word in a specific document leads to the respective word count vector. Thus each image is initially represented by a vector of length M, if M denotes the size of the vocabulary.

Employing these representations, we estimate the model parameters of a CTM based on a training set of images out of all images in the database. Since we are interested in an image representation suited for retrieval in large databases, we utilize the learned model and infer the topic proportions for each image in the dataset. This gives a higher level description of the image content. In the following we represent the images by these low-dimensional topic vectors.

This image representation then allows us to perform image retrieval by comparing the representations of two images based on some similarity measures and keeping the image which are most similar to some query image. In our experiment we employ four different similarity measures to compare two images' representations. The three commonly known standard approaches, cosine similarity, L1 distance and the Jenson-Shannon divergence are supplemented by a fourth similarity measure adopted from language-based information retrieval, which measures the likelihood that the model of a document generated the query image, which is described in detail in [4] and [7]. Note that the first three measures are only based on the topic vectors of the respective images where the last measure uses a combination of the image model from the CTM and the bag-of-words model.

3. DATABASE

All experiments are performed on a database consisting of approximately 246,000 images. The images were selected from all public Flickr images uploaded prior to Sep. 2006 and labeled as *geotagged* together with one of the following tags: *sanfancisco*, *beach*, and *tokyo*. Of these images only images having at least one of the following tags were kept: *wildlife, animal, animals, cat, cats, dog, dogs, bird, birds, flower, flowers, graffiti, sign, signs, surf, surfing, night, food,*

Cat.	OR list of tags	#
1	wildlife animal animals cat cats	28,509
2	dog dogs	24,660
3	bird birds	20,908
4	flower flowers	25,457
5	graffiti	21,888
6	sign signs	14,333
7	surf surfing	29,552
8	night	33,142
9	food	18,602
10	building buildings	16,826
11	goldengate goldengatebridge	23,803
12	baseball	12,372
	Total # of images	
	Note: Images may have multiple tags	246,348

Table 1: List of categories, their corresponding tag set, and their number of images.

building, buildings, goldengate, goldengatebridge, baseball. The images can thus be grouped into 12 categories as shown in Table 1. Note that no preprocessing was applied to these images and the database was automatically generated without manually cleaning out any images.

4. EXPERIMENTAL RESULTS

We evaluate the performance of the described approach in a query-by-example retrieval task, i.e., given a query image the goal is to find images of similar content in the database. For each category five query images are randomly selected from the database, resulting in an overall count of 60 test images. Having computed the L most similar images to each query image, we then rated the performance of our models by means of user studies: Users were presented the retrieved images and asked to count the number of correctly retrieved images for each method. The final score is then computed as the average score over all images and users. Note that the judgment of the users is subjective, as each user may perceive the content of an image slightly differently. Thus we also compute the standard deviation from the average score.

For our experiments we constructed a visual vocabulary out of 12 randomly selected, non-overlapping subsets of all visual words of the image database. Each subset contained 500,000 visual words and was clustered to produce 200 distinct visual words. The clusters were then merged, resulting in an overall vocabulary size of 2,400.

4.1 Parameter Estimation

Since it is not obvious how to choose the number of topics K in the CTM and also the number of images in our training set, we first determine suitable values for these learning parameters. To measure the performance of the model with respect to these parameters, we apply the commonly used *perplexity* measure [2]. The perplexity indicates how good the model is able to generalize on held out data and decreases monotonically in the likelihood of this data. Thus lower perplexity values implicate a better model performance. The
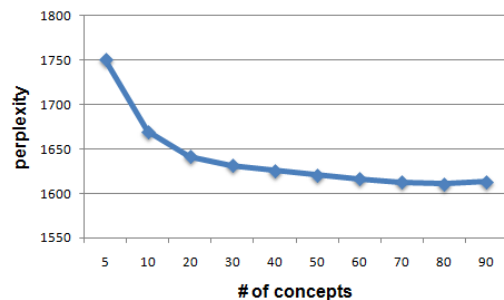

Figure 2: Number of concepts in the CTM plotted against the perplexity of a previously unseen test set.

perplexity on a held out dataset D_{test} is defined by [2]:

$$per(D_{test}) = \exp\left\{-\frac{\sum_{d=1}^{D} \log p(\mathbf{w_d})}{\sum_{d=1}^{D} N_d}\right\} \qquad (2)$$

In our first experiment we evaluate the influence of the choice of the number of topics K on the performance of the model. We train a CTM on a subset of 50,000 images with a varying number of topics. The perplexity is calculated on a previously unseen test set of 25,000 images and is shown in Figure 2. A small number of topics lead to a high perplexity, since the model fails to fit to the complex test data. However, for large K the perplexity is rather constant. Since the minimum perplexity is reached at 80 concepts, the number of concepts would usually be set to 80. However we also need to consider the dimensionality of our model where a smaller number of topics is preferred in large-scale database to represent the images. Observing that the difference in perplexity values is rather small between 80 topics and 50 topics, we choose to set $K = 50$ in all our subsequent experiments.

Figure 3 shows the calculated perplexity plotted against the number of training images used to train a CTM consisting of 50 topics. In contrast to our previous experiment, the perplexity does not seem to follow a clear pattern. Moreover the dependence of the perplexity on the number of training samples does not seem to be as pronounced as is was for the number of topics; the range of perplexity values is rather small compared to the values obtained by changing the number of concepts. These results suggest that it is not necessary to train the model on the whole corpus, which is a big advantage when using a very large database like ours.

4.2 Similarity Measures

In order to evaluate the different similarity measures on the image retrieval task a CTM is trained using a training set of 20,000 images and $K = 50^1$ concepts. Representing all images in the database by the CTM model and applying the four different similarity measures described in Section 2.2 we find the 19 closest images of the database to each query image. Figure 4 shows examples of query images and their four most similar images retrieved.

In a user study the retrieved images were presented to eight

---
[1]The parameter λ of the IR distance measure has been set to 0.8 throughout the experiments
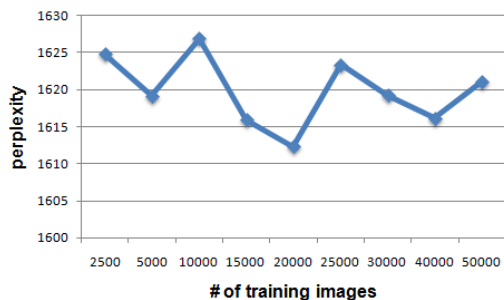
Figure 3: Number of training images plotted against the calculated perplexity on a previously unseen test set with the number of topics set to $K = 50$.
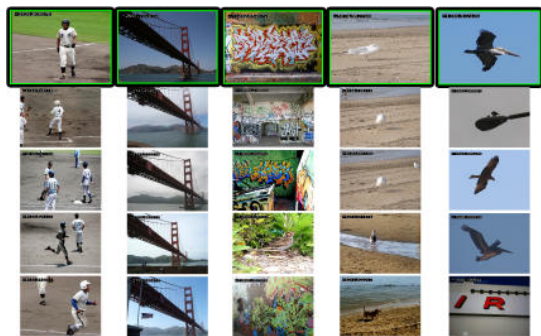

Figure 4: Examples of retrieval results. The first image of each column shows the query image, followed by the four most similar images retrieved.

Figure 5: The average number of correctly retrieved images for a CTM based image representation and different similarity measures.

Figure 6: The average number of correctly retrieved images of the CTM-based image representation compared to other approaches. All approaches are using the IR measure.

test users who were asked to count the number of correctly retrieved images for each measure. The average number of correctly retrieved images of each similarity measure according to the users' judgment is depicted in Figure 5, the vertical bars represent the standard deviation. Clearly, the probability measure adopted from information retrieval [7][4] outperforms all other similarity measures. Between all measures that solely use the topic proportions to measure the distances, the Jensen-Shannon divergence returns the best result, followed by the L1 distance. Notice that these results are consistent with the results of our previous work [4].

4.3 CTM vs. LDA and pLSA

In our last experiment we compare the presented approach to previous approaches [4][5]. Using the same setup as before, users were presented retrieved images of three models: the CTM, LDA and pLSA combined with their respective best performing similarity measure, which is in all cases the measure adopted from IR (see [4]). The results of this user study are displayed in Figure 6, showing the mean and standard deviation of the number of correctly retrieved images. It can be seen that the average number of correctly retrieved images of the CTM-based representation is lower compared to the score of the LDA- and even that of the pLSA-based description. This result is surprising as, when applied to text documents, the CTM has been shown to produce decent results [1]. The inferior performance of the CTM model in our database might be due to the number of topics in the
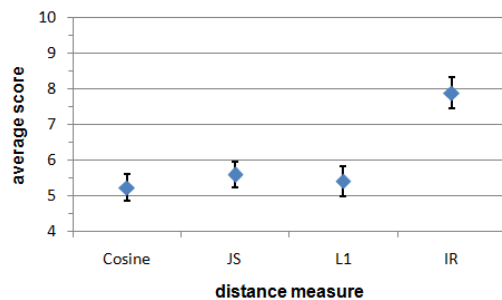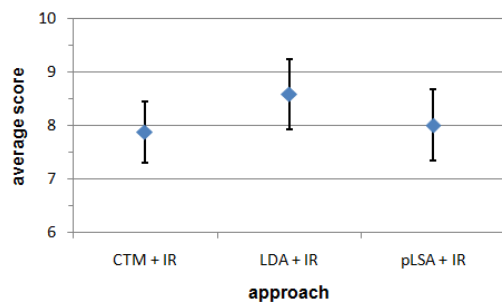
model. As the database is quite noisy, the number of topics might have been chosen to small to allow for dependencies between the topics. However, a large number of topics contradicts with the aim of finding a suitable lower dimensional representation that allows fast retrieval in large databases. This issue needs to be addressed in further research.

5. CONCLUSIONS

We have studied the representation of images by the correlated topic model and evaluated its performance by means of a large-scale image retrieval task. The CTM models the correlation between different topics and is therefore an improved generative model compared to the previously applied LDA. However, the results of our experimental evaluation have shown that the model does not perform superior to previous approaches. Future work will consist of examining the CTM in more detail in the context of image retrieval, especially by validating our results for a larger number of topics.

6. REFERENCES

[1] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35, 2007.
[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
[3] T. Hofmann. Unsupervised learning by probabilistic

latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[4] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 17–24, New York, NY, USA, 2007. ACM.

[5] R. Lienhart and M. Slaney. Plsa on large scale image databases. *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, 4, 2007.

[6] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[7] X. Wei and W. B. Croft. Lda-based document models for ad-hoc retrieval. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185, New York, NY, USA, 2006. ACM.