Exploring Scale-Induced Feature Hierarchies in Natural Images

Jukka Perkiö
Helsinki Institute for Information Technology
Finland
jperkio@cs.helsinki.fi

Tinne Tuytelaars
Katholieke Universiteit Leuven
Belgium
Tinne.Tuytelaars@esat.kuleuven.be

Wray Buntine
NICTA
Australia
wray.buntine@nicta.com.au

Abstract

Recently there has been considerable interest in topic models based on the bag-of-features representation of images. The strong independence assumption inherent in the bag-of-features representation is not realistic however: patches often overlap and share underlying image structures. Moreover, important information with respect to relative scales of the features is completely ignored, for the sake of scale invariance. Considering both spatial and scale-based constraints one can derive spatially constrained natural feature hierarchies within images. We explore the use of topic models that build such spatially constrained scale-induced hierarchies of the features in an unsupervised fashion.

Our model uses standard topic models as a starting point. We then incorporate information about the hierarchical and spatial relations of the features into the model. We illustrate the hierarchical nature of the resulting models using datasets of natural images, including the MSRC2 dataset as well as a challenging set of images of trees collected from the Internet.

1 Introduction

Recently there has been considerable interest in using statistical topic models such as PLSA [1] or LDA [2] on images in the same way as they have traditionally been used in (text) information retrieval. Some noteworthy examples include [8, 4, 5]. To make these models applicable to visual data, a visual equivalent of the concept of words in a text document is needed. To this end, one typically samples patches from the images, either using an interest point detector or sampling patches on a regular grid. These patches are then represented using a robust descriptor like SIFT [6] and vector quantized to generate a *visual vocabulary*, through which the images are represented. A very common representation is the *bag-of-features* representation, which only takes into account the counts of visual words within an image, hence the bag-of-features name.

Visual words in images do not behave the same way as words in a text document though. As a result, directly translating the topic models developed for text to the image domain may not give the best results. Especially when going for a fully unsupervised approach, reported results on visual object discovery based on topic models, as in [4], have been limited to artificially constructed databases with a limited number of objects that can easily be distinguished. In this paper, we explore an extension of those standard statistical topic models, that takes some image-specific characteristics into account.

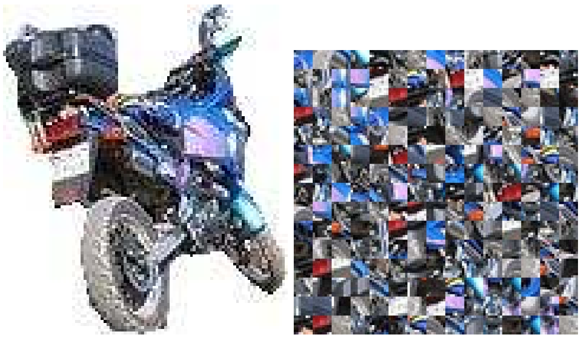

Figure 1. Without spatial configurations between the visual words, a lot of the semantics of the image is gone (image courtesy of John Winn).

In a text document, it is reasonable to assume that given a topic words are independent, based on the notion of exchangeability. For images, on the other hand, patches typically overlap and share the same underlying spatial structure, which typically extends over a larger area of the image. When scrambling the words of a text, one might still guess the topic. When scrambling the patches of an image, on the other hand, the meaning is most often lost (see figure 1). Neither exchangeability nor independence given the topic seem to hold. Hence, we want to explore an alternative topic model, that captures the dependency between overlapping patches and indirectly encodes the spatial structure of the image.

Another important aspect when dealing with images as opposed to text documents is the notion of *scale*. Characteristic features in an image can be found over a range of scales. This is the result of a complex interaction between the distance between the camera and the object on the one hand and the intrinsic scale of a feature (e.g. a wheel versus a car) on the other hand. Scale space theory [7] has taught us that we should hence analyze an image over different scales, e.g. using a Gaussian image pyramid.

For sake of simplicity though, practical implementations take an opportunistic approach. Some of them (e.g. [9]) extract features only at a single scale, which gives good results on datasets where significant scale changes are rare. Others do indeed extract features from scale space (either in a scale invariant way or using a multi-scale approach). However, to make the methods invariant to scale changes, all scale-related information is then filtered out. In a bag-of-features representation, the visual word ID only encodes the appearance of the image patch, not its scale. This ensures that a similar representation is found when the distance between camera and object changes, yet at the same time throws away a lot of relevant information, e.g. on relative scales between two different features.

Objects contain different characteristic features at different scales. For instance, from a distance (at large scales), grass or trees may look similar as they both are more or less homogeneous and green. When zooming in though (looking at finer scales), the structures of the individual leaves and branches show up. Rather than modeling the visual concept of grass or trees as merely a collection of homogeneous and textured patterns, it would be interesting to encode that the homogeneous patterns are found at a larger scale, with the textured patterns at finer scales. This can be done by conditioning the topics of the smaller scale patches on the topics of the larger scale patches with which they overlap. With such a hierarchical model, the larger scale, homogeneous green patches could then correctly be assigned to either the ’grass’ or ’trees’ topic.

Assuming a grid-like sampling strategy over different scales, a bag-of-features representation for images, and a statistical topic model based on the visual vocabulary, one can try to remedy the rather unrealistic independence assumption inherent in the bag-of-features representation. Since both scale- and spatial information are captured in the sampling, it is natural to try to use this information in the final model. One way of doing this is to consider the sampling patterns within individual images as a graph defined by the scale and the overlapping relations between patches. For each image one can define a directed graph so that the patches of the largest scale are root nodes and they have overlapping smaller scale patches as children, each level of the graph being of a different scale. This way we have for each image a directed graph that contains information about both the scale and the spatial structure of the image. Each node may have several parents and several children, and the graph is spatially constrained based on the sampling strategy (see also figure 2).

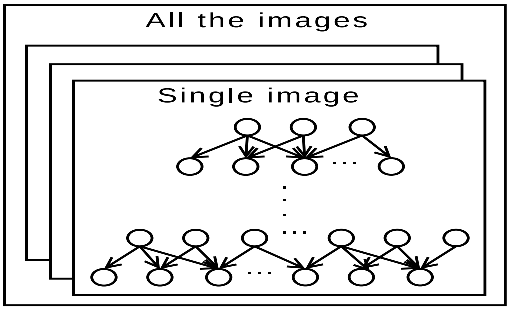

Figure 2. Overlapping samples produce the sampling graph that is used for defining the dependencies between image patches.

Related Work Other spatial or hierarchical extensions to the standard statistical topic models have been proposed in the literature before, albeit mostly in a (semi-)supervised setting and/or not dealing with the scale issue. Verbeek et al. [9] have proposed the Markov Field Aspect Model, which combines the local smoothness of Markov Random Field models with the global consistency of PLSA. In fact, their model is very similar to ours, except that they condition the topic for a visual word on the topics of the neighboring visual words, whereas we condition it on the topics of overlapping patches found at a different scale. By doing so, we bring in ideas from scale space theory and make the model robust to scale changes. Simultaneously, we believe that looking at overlapping patches is a more natural choice than selecting 4-neighbors. Moreover, they do not report any results in an unsupervised setting.

Probably most similar to our work is the hierarchical Dirichlet Process - hidden Markov tree proposed by Kivinen et al. [3]. They use a tree-structured graphical model, incorporating the dependencies between latent topics at nearby locations and scales. However, they only show results in a supervised setting, and do not explore the learnt hierarchy.

Other hierarchical extensions of statistical topic models, such as hierarchical LDA, have been applied to images as well [10, 11]. However, these focus on finding hierarchies of object categories, whereas our model focusses on the hierarchical structure within a single object category.

2 Methodology

We start by sampling overlapping square image patches using a grid-like sampling pattern. As a result we have N image patches of S different sizes from I images. For each patch we then apply an image feature extractor (for the experiments in this paper we use the SIFT [6] features) to derive a vector representation for patches after which we produce a visual vocabulary clustering vectors to L clusters using k–means, as is customary in the image processing community. Each cluster now represents one visual word and each image is represented using bag–of–words representation with the obtained visual vocabulary. Because of the grid like sampling of image patches at different scales, we have for each image a directed graph G_i that gives us the spatial dependencies between the overlapping image patches. The largest scale patches are the root nodes and the smallest scale patches are the leaves. Table 1 gives a summary of the notation used in the following and throughout this paper.

Symbol	Explanation
N	Number of image patches.
S	Number of different scales for image patch extraction.
I	Number of images.
L	Size of visual vocabulary.
G_i	Graph of patch dependencies for image i.
K	Number of topics.
θ	Multinomial parameters, i.e. visual word probabilities for topics as row vectors.
α	The dependency structure between topics.
$\mathfrak{p}_n$	Image patch n.
w_n	Visual word for patch n.
t_n	Topic for patch n.
$\mathcal{P}(t_k)$	Parents of topic k.
Z	Normalizing constant.

Table 1. The notation used throughout the paper.

We want to estimate a K topic hierarchical topic model in which the hierarchy is based on the spatial dependencies between individual image patches. For that we assume following:

- Each topic is represented by a vector θ_k so that $\sum_i \theta_k^i = 1$. In other words we have a simple mixture of multinomials. Here we are only interested in the visual word / topic probabilities $p(\mathfrak{p}_n|t_n, \theta)$, not in the mixing proportions. Initial estimation of these parameters is not critical.

- In the graph $G = \cup_i G_i$ each patch (visual word) is assigned to a single most probable topic for that patch. This way we get a hierarchical representation of topics, which is not optimal yet.

- The topic dependencies are modelled by a matrix α and the topic probabilities given topics' parents are given by $p(t_n|\mathcal{P}(t_n), \alpha)$, which is in the exponential family.

Given the above, we want to maximize the likelihood of topic assignments and parameters θ and α.

If there were no dependencies between patches (visual words), the likelihood of topic assignments and parameters θ would be maximized simply by some non-hierarchical topic model, e.g. LDA or simple mixture of multinomials. However, we have the hierarchical dependencies given by the graphs G_i and modelled by the parameters α and the estimation process is more complex. The likelihood we are interested in is

$$p(\mathfrak{p}, t|\theta, \alpha) = \prod_n p(\mathfrak{p}_n|t_n, \theta) p(t_n|\mathcal{P}(t_n), \alpha), \quad (1)$$

where

$$p(\mathfrak{p}_n|t_n, \theta) = \theta_{t_n, w_n}, \quad (2)$$

and

$$p(t_n|\mathcal{P}(t_n), \alpha) = \frac{1}{Z} \exp \sum_{t_i \in \mathcal{P}(t_n)} \alpha_{t_n, t_i}. \quad (3)$$

The above likelihood can be maximized using an iterative procedure: The topic assignments are sampled in the graphs G_i, parameters θ are updated using EM and parameters α are optimized using stochastic gradient descent. Initially the parameters θ can be estimated using non-hierarchical topic models, e.g. LDA, mixture of multinomials, etc. The exact estimation procedure is as follows:

1. Estimate initial parameters θ using some discrete topic model, e.g. LDA, mixture of multinomials, etc.

2. Set the topic assignments according to the parameters θ.

3. Optimize parameters α.

4. Sample the graphs G_i for topic assignments.

5. Update the topic assignments to the ones that were sampled the most.

6. Optimize parameters θ.
7. Update topic assignments according to the parameters θ.
8. Optimize parameters α.
9. Repeat from 4.

The above procedure produces new estimates for parameters θ, α and for the topic assignments in the graphs G_i, that take into account the hierarchical dependencies between visual words because of the dependencies between image patches. After we have estimated the topic assignments and the parameters α we can use either of them to define a topic hierarchy. The resulting hierarchy is not a hierarchy in a strict sense, since it may contain cyclical dependencies. This is because the estimation is done in an unsupervised fashion and the same visual word may be both the parent and child of itself, since they are ambiguous in relation to image patches.

3 Experiments

We use two different datasets for the experiments that are detailed below:

1. Trees dataset: 290 images of trees collected from the Internet.
2. MSRC2 dataset: 591 images from Microsoft research Object recognition database[1].

For each set above the images are converted to grayscale and scaled to contain 150000 pixels retaining the aspect ratio. The patches are sampled at two scales per octave from 27 pixels to 288 pixels. We use the SIFT feature detector [6] for each patch in order to derive 128-dimensional vectors that are clustered to produce the final visual vocabulary of size 500. Table 2 summarizes the experimental setting.

	Trees	MSRC
Size of visual vocabulary	500	500
Number of images	290	591
Number of patches	1155240	2330008
Number of scales	8	8
Number of topics	10, 15, 20	10, 15, 20

Table 2. Basic parameters for the experimental models.

The performance of the proposed method can be evaluated both through the likelihood scores and also through visual examination. We use LDA to estimate the initial parameters θ. Table 3 shows BIC and AIC scores for six flat and hierarchical models of different size on the dataset. It can be seen that both scores are better for the hierarchical version of the model, but the difference is not huge. That is quite understandable, since the flat LDA models are quite optimal already and the incorporated hierarchical dependencies do not bring that much extra information. On the other hand the improvement, even if small, is clear in each case we tried. For the trees dataset two bigger models benefited the most from using the hierarchical model, whereas for the MSRC2 dataset the smallest model benefited the most.

	BIC			AIC		
Model	flat	hier.	%	flat	hier.	%
T 20	9.889	9.166	7.3	9.769	9.042	7.4
T 15	10.340	9.580	7.4	10.251	9.488	7.4
T 10	10.771	10.275	4.6	10.712	10.214	4.6
M 20	17.900	17.593	1.7	17.774	17.461	1.8
M 15	18.596	18.317	1.5	18.501	18.219	1.5
M 10	20.326	18.988	6.6	20.262	18.924	6.6

Table 3. BIC and AIC scores improve when using the hierarchical model. Values are divided by 10^6. Trees dataset is denoted with T and MSRC2 dataset with M. The improvement for both scores is shown in percents in the fourth and seventh column of the table.

Because of space constraints we only show visual results for the 20 topic model of the trees dataset. Table 4 shows an example patch for the five most important visual words for each topic in that model. One has to note though that because of the ambiguity of visual words vs. image patches visualizing them is not an easy task. However the table gives a general feeling about the nature of and the differences between the topics. Figure 3 shows the topic hierarchy for the same model. We have omitted the very weak dependencies between the topics. As it was noted it is possible to have cyclical dependencies between same and different topics. We only show cyclical dependencies between different topics. Below we will discuss shortly some of the topics and hierarchy shown in Table 4 and Figure 3.

Topic 15 has topics 1, 9 and 14 as its children. Topic 15 is for larger scale patches, the patches that are shown contain many details that cannot be seen in the image. Most importantly the topic 1 is a perfect subset of the textures represented in topic 15, whereas topic 14 is finer scale details of leaves and branches also present in topic 15. It has also defining feature of having many patches that are easily partitioned to two distinct areas. Topic 9 is of larger scale again but mostly related to finer textures.

Topic 10 has topics 3 and 16 as its children. Topic 10 is mostly related to rather clear detailed leaves and branches. Patches are of varying scale but on the bigger side. Topic

[1] http://research.microsoft.com/en-us/projects/objectclassrecognition/

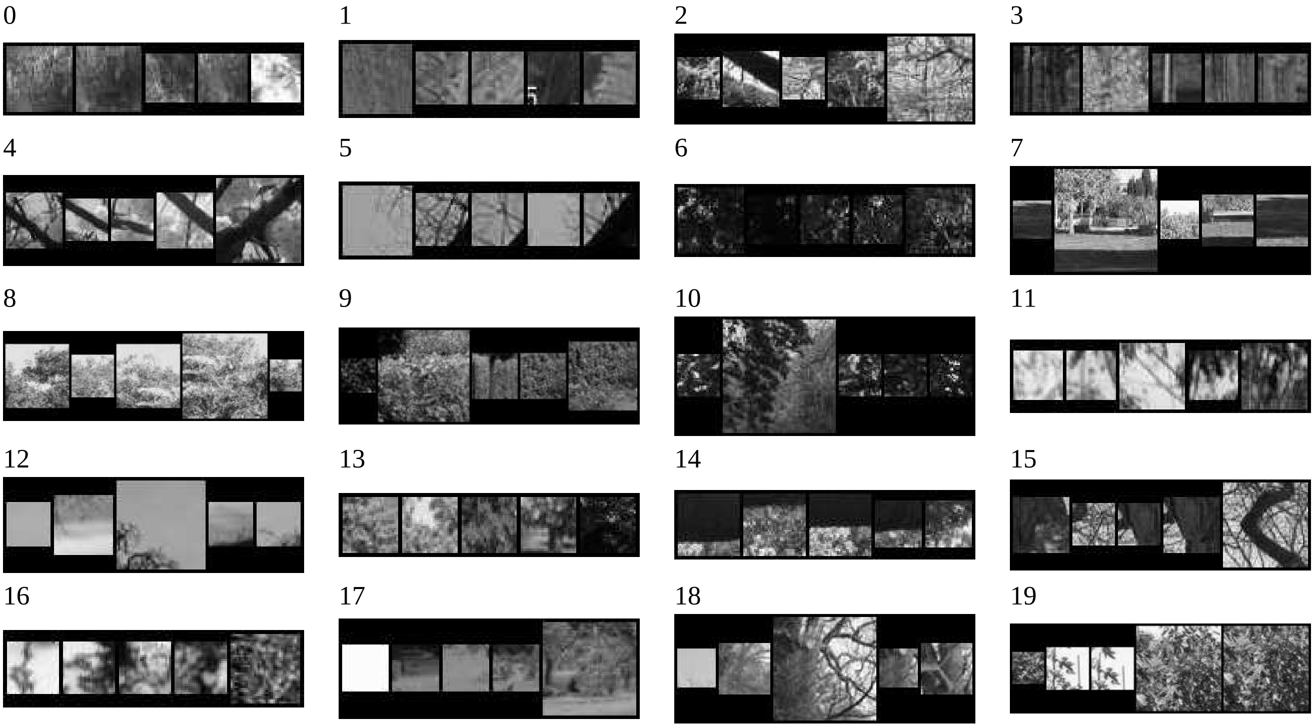

Table 4. For each topic of the 20 topic model of trees dataset 5 most probable visual words are shown. Since visual words are ambiguous, i.e. one word maps to several image patches, visualizing them is hard. One should see much more image patches than it is possible to show due to space constraints. One should also note that in the table there are patches in all scales ($27\times27 \rightarrow 288\times288$ pixels). This still makes the visual inspection harder. The patches above are shown from randomly selected images.

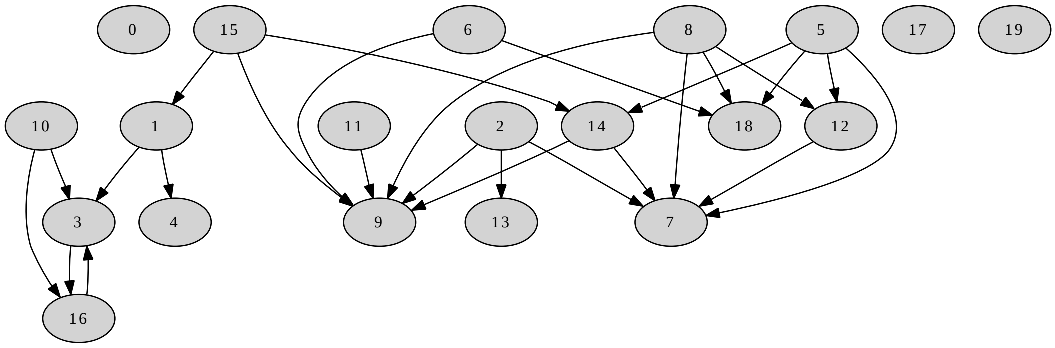

Figure 3. The topic graph for the 20 topic model for the trees dataset. One should see the Table 4 for information about the topics.

16 is for smaller fine detail that can be found in patches of topic 10. Topic 3 is smaller scale topic for vertical trunks of trees. Topic 3 and 16 have cyclical dependency and because of this it is more understandable why topic 3 is also a child of topic 10.

Topic 14 has topics 7 and 9 as children. Topic 7 is similar to 14, since both of them have patches of clearly bipartitionable nature. Topic 9 has textures that are also present in topic 14.

Topic 9 is a special case in a sense that it is a child of many topics. This is understandable since, the textures present in that topic are easily found in many patches of other topics.

Topic 7 is also a child to a number of other topics. The nature of topic 7 is mostly related to larger, clearly distinguishable areas in the patches.

It is also interesting to look at dependencies of higher depth than two. Such is e.g. topic chain $5 \rightarrow 14 \rightarrow 9$ Topic 5 has rather smooth and large surfaces , topic 14 has more textured surfaces and topic 9 is mostly concerned in rather fine texture. Another example is the chain $15 \rightarrow 1 \rightarrow 3 \rightarrow 16$. Topic 15 has even and coarser textures, topic 1 has the coarser texture present, topic 3 have slightly different type of coarser texture and topic 16 have very fine detail in small scale patches that describe the coarser textures.

One can conclude that the hierarchies do make sense even though the visualizations are quite demanding due to the ambiguity of image patches vs. visual words.

4 Discussion

We have proposed a statistical topic model for images that takes the dependencies between overlapping image patches into account.

It properly deals with scale: rather than simply making everything scale-invariant and forgetting about scale from that point onwards, we take into account that objects typically show different structures at different scales. We go back to the basic idea of scale space: structures in an image exist over a range of scales. We link these structures found at different scales, taking their relative scales into account, resulting in a more discriminative, yet still scale-invariant model.

Linking overlapping patches at different scales is more natural than linking neighbors, as was done in [9]. Yet, apart from being able to represent structures at different scales, it also has the same smoothing effect, as neighboring patches share many of their parents.

Because we condition topics on the topics of overlapping larger-scale patches, we get a topic-hierarchy. Unlike results obtained with hierarchical LDA etc., our hierarchy focusses on the underlying features, at different scales, from which objects are composed, rather than hyponyms/hypernyms.

For the experiments in this paper we have used only SIFT features but using other types of features would most likely improve the models further. Combining SIFT, color and specific texture features seem a natural way to continue. In that kind of setting the scale would be handled by the sift features and the texture and color would function as discriminative power within different scales. On the other hand using an extensive number of features also adds the computational cost.

5 Conclusions

We have proposed a new statistical topic model for images, that allows us to find topics arranged in a scale-induced hierarchy, with the higher level topics describing large scale events and the lower level topics focussing more on the details. This model naturally takes the scale of the features into account, via the hierarchical structure of the topics, yet it remains scale-invariant (as opposed to other schemes that e.g. directly include the scale in the descriptor or build different bag-of-features for different scale intervals). Finally, by including dependencies between overlapping patches, we get smoother results than what is typically obtained with a flat LDA.

References

[1] Thomas Hofmann, Probabilistic Latent Semantic Indexing , SIGIR, 1999.

[2] D. Blei, A. Ng, and M. Jordan, Latent Dirichlet allocation, Journal of Machine Learning Research, 3:9931022, 2003.

[3] J.J. Kivinen, E.B.Sudderth, and M.I. Jordan, Learning Multiscale Representations of Natural Scenes Using Dirichlet Processes,ICCV 2007.

[4] J. Sivic, B.C. Russell, A.A. Efros, A. Zisserman, and W.T Freeman, Discovering Objects and their Localization in Images, ICCV, pp.370-377, 2005.

[5] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica Perez, D. and T. Tuytelaars, A Thousand Words in a Scene, PAMI, 29(9):1575-1589, 2007.

[6] D. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, IJCV, 60(2):91-110, 2004.

[7] T. Lindeberg, Scale-Space Theory in Computer Vision, Kluwer, 1993.

[8] L. FeiFei and P. Perona, A Bayesian Hierarchical Model for Learning Natural Scene Categories, CVPR, pp. 524-531, 2005.

[9] J. Verbeek and B. Triggs, Region Classification with Markov Field Aspect Models, CVPR, 2007.

[10] J. Sivic, B.C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, Unsupervised discovery of visual object class hierarchies, CVPR, 2008.

[11] E. Bart, I. Porteous, P. Perona, and M. Welling, Unsupervised Learning of Visual Taxonomies, CVPR, 2008.

[12] B. Leibe, A. Leonardis, and B. Schiele, Robust Object Detection with Interleaved Categorization and Segmentation, IJCV, 77(1-3):259-289, 2008.

[13] P.F. Felzenszwalb and D. P. Huttenlocher, Pictorial Structures for Object Recognition, IJCV, 61(1):55-79, 2005.