Decomposition, Discovery and Detection of Visual Categories Using Topic Models

Mario Fritz and Bernt Schiele
Comuter Science Department, TU-Darmstadt, Germany
{fritz,schiele}@cs.tu-darmstadt.de

Abstract

We present a novel method for the discovery and detection of visual object categories based on decompositions using topic models. The approach is capable of learning a compact and low dimensional representation for multiple visual categories from multiple view points without labeling of the training instances. The learnt object components range from local structures over line segments to global silhouette-like descriptions. This representation can be used to discover object categories in a totally unsupervised fashion. Furthermore we employ the representation as the basis for building a supervised multi-category detection system making efficient use of training examples and outperforming pure features-based representations. The proposed speed-ups make the system scale to large databases. Experiments on three databases show that the approach improves the state-of-the-art in unsupervised learning as well as supervised detection. In particular we improve the state-of-the-art on the challenging PASCAL'06 multi-class detection tasks for several categories.

1. Introduction

Object representations for categorization tasks should be applicable for a wide range of objects, scaleable to handle large numbers of object classes, and at the same time learnable from a few training samples. While such a scalable representation is still illusive today, it has been argued that such a representation should have at least the following properties: it should enable sharing of features [27], it should combine generative models with discriminative models [13, 10] and it should combine both local and global as well as appearance- and shape-based features [16]. Additionally, we argue that such object representations should be applicable both for unsupervised learning (e.g. visual object discovery) as well as supervised training (e.g. object detection).

The main focus of this paper is therefore a new object representation that aims to combine the above mentioned properties to make a step towards more scalable object representations applicable to a wide range of objects and suited both for unsupervised as well as supervised learning. Therefore, the first main contribution of this paper is a novel approach that allows to learn a low-dimensional representation of object classes by building a generative decomposition of objects. These learned decompositions of objects contain both local appearance features as well as global silhouette features shared across object classes. This generative model of objects is directly applicable to unsupervised learning tasks such as visual object class discovery. The second main contribution of the paper is then to combine the low-dimensional and generative decomposition of objects with a discriminative learning framework to enable supervised training and competitive object class detection. The third contribution of the paper is a series of experiments which show the properties of the approach (local vs. global features, feature sharing, unsupervised vs. supervised learning) and compares the approach with the state-of-the-art. Interestingly, the approach outperforms both unsupervised techniques as well as supervised techniques on various tasks on common databases.

The paper is structured as follows. In Section 2 we describe how the generative decomposition is learned from data. The obtained representation is used in Section 3 for unsupervised learning problems whereas Section 4 builds a full object class detection system on top of it. Finally Section 5 provides further quantitative evaluations of the model and a comparison to the state-of-the-art on the challenging PASCAL'06 database as well as a shape database.

Related Work Feature representations based on gradient histograms have been popular and highly successful ranging from local statistics like SIFT [18], over part-like fractions [14] to the representation of entire objects [6, 1]. Based on their success we build our method on a dense grid of local gradient histograms inspired by [6].

In terms of generative modeling, we build on the success of topic models (e.g. [12, 2, 11]). They have gained increasing attention in computer vision ranging from unsupervised category discovery [24, 17, 3], over classification [22, 15] to detection [26, 8, 1]. Often local feature representations

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

are employed [24, 15] that neglect the spatial layout with a few exceptions such as [8, 26, 1]. In contrast we employ a dense representation based on gradient histograms that explicitly retains the spatial feature layout. Topic model learning is then employed to decompose objects into constituent parts in an unsupervised fashion. Thereby a versatile multiclass object representation is derived. Without posing any constraints on the locality of the topics, we obtain topics ranging from global silhouette types to local edge features.

2. Decomposition of Visual Categories

In this section we describe our approach to decomposition of multiple visual categories by combining dense gradient representations and topic models. Starting from the image, we first present our data representation. Then we describe how we apply the topic model to this representation and provide visualizations and insights for the obtained model as well as a quantitative evaluation on an unsupervised learning task.

2.1. Data Representation

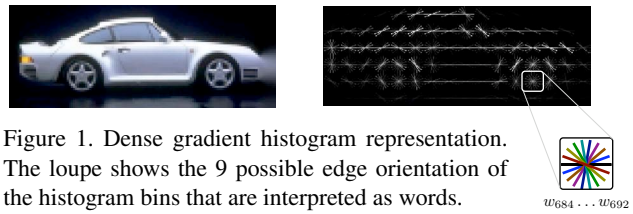


Figure 1. Dense gradient histogram representation. The loupe shows the 9 possible edge orientation of the histogram bins that are interpreted as words.

Inspired by [6], we compute gradients on each color channel of the images and use the maximum response to obtain a grid of histograms that overlays the image. Each histogram in the grid has 9 orientation bins equally spaced from 0° to 180° to represent the unsigned gradient orientation. An example of such an encoding is visualized in Figure 2.1. In each cell, the 9 possible edge orientations associated with the orientation bins are displayed by short lines. The grayscale value encodes the accumulated gradient magnitude in each bin. The size of the cells in the grid is 8×8 pixels.

As the following topic models operate on discrete word counts, we normalize the histograms to have a constant sum of discrete entries. We decided not to compute a redundant coding like the blocks in the HOG descriptor [6] as we believe that the introduced non-linearities by local normalization would hinder the fitting of the probabilistic model.

2.2. Topic Models

To define a generative process for our data representation, we employ probabilistic topic models [12, 2, 11] which were originally motivated in the context of text analysis. As

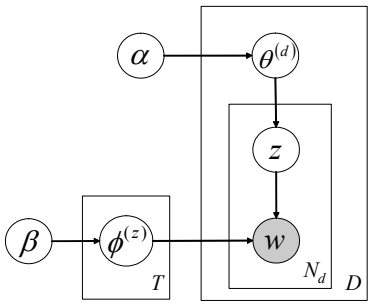


Figure 2. LDA model as formulated by [11].

it is common habit we adopt the terminology of this domain. In the following, a document d refers to a sequence of words $(w_1, w_2, \dots, w_{N_d})$, where each w_i is one word occurrence. The underlying idea of these models is to regard each document as a mixture of topics. This means that each word w_i of the total N_d words in document d is generated by first sampling a topic z_i from a multinomial topic distribution $P(z)$ and then sampling a word from a multinomial topic-word distribution $P(w|z)$. Therefore the word probabilities for the combined model are:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z_i = j)P(z_i = j) \tag{1}$$

where T is the number of topics and $P(w_i|z_i = j)$ as well as $P(z_i = j)$ are unobserved . According to the notation of [11], we will abbreviate

$\theta^{(d)}$: topic distribution $P(z)$ for document d
$\phi^{(j)}$: topic-word distribution $P(w_i|z = j)$ for topic j

The particular topic models differ on the one hand in which additional hyperparameters/priors they introduce and on the other hand in how inference and parameter estimation is performed. We will discuss the *Latent Dirichlet Allocation* model [2] in some more detail focusing on the version presented in [11] that uses Gibbs sampling for inference and estimation. The graphical representation of this model is depicted in Figure 2. It visualizes the process that generates a total of D documents d, where each document has N_d words. Above we already described how each word w_i of a particular document is generated. In the full model, there are 2 additional hyperparameters, α and β, which place symmetric dirichlet priors on the topic distribution of each document $\theta^{(d)}$ and the topic-word distributions $\phi^{(j)}$ respectively. As the setting for α and β is common to all documents, these act as forces that impose global tendencies on these distributions. Intuitively, the prior α for the topic distribution θ favors co-activation (sharing) of multiple topics for each document for values

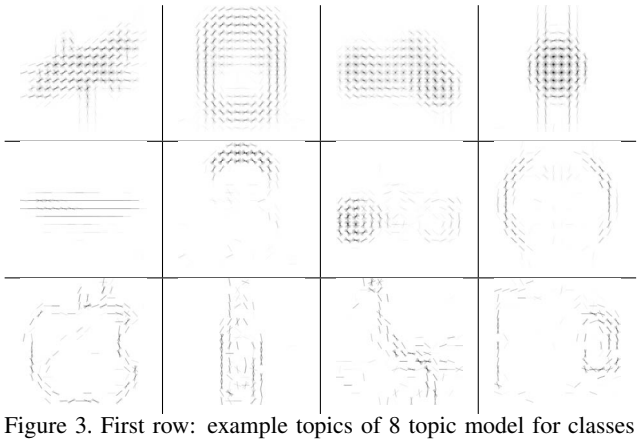


Figure 3. First row: example topics of 8 topic model for classes airplane, face, motorbike, watch. Second row: example topics of 50 topic model for the same classes. Third row: example topics of 100 topic model jointly learned on apple-logos, bottles, giraffes, mugs and swans.

larger than 1, whereas smaller values result in sparser topic distribution - ultimately having single topics explaining whole documents (clustering). Consequently, the sparseness of the topic-word distribution $\phi^{(j)}$ is affected by this choice. The second parameter β, has a direct smoothing effect on the topic distributions.

For more details on the models, inference and estimation, we refer to [2] and [25]. The idea behind the employed Gibbs sampling procedure is that all topic assignments z_i are initialized (typically randomly) and then iteratively updated in a random order. To perform such a single update, a topic is drawn from the conditional distribution $P(z_i|\Omega \setminus z_i)$ and assigned to z_i, where $\Omega \setminus z_i$ denotes all observed and unobserved variables but z_i. This is repeated for a fixed number of iterations.

3. Discovery of Visual Categories

In this section we describe how the representation from Section 2.1 is linked to the generative model from Section 2.2 and perform a quantitative evaluation on an unsupervised learning task.

We use the orientation bins of the histograms described in Section 2.1 as word vocabulary in Section 2.2. For histograms computed on a m by n grid with b bins for each orientation histogram, our vocabulary is of size $|V| = m \cdot n \cdot b$. As each word is associated with a gradient orientation at a grid location, this representation preserves quantized spatial information of the original gradients. The topic model is trained on the documents given the encoded training examples. The representations that we promote are given by the topic activations of the document in the latent space.

To prove the effectiveness of our representations and to compare our work with previous approaches we first present quantitative results on an unsupervised ranking task [8] and then provide further insights connected to the multi-class data we use in Section 5.3.

	airplane	cars rear	face	guitar	leopard	motorbike	wrist watch	average
out method	100%	83%	100 %	91%	65%	97%	100%	91%
Fergus [8]	57 %	77%	82%	50%	59%	72%	88%	69%
Schroff [23]	35%	–	–	29%	50%	63%	93%	54%

Table 1. Comparison to other approaches on re-ranking task of google images. Performance is measured in precision at 15% recall. In contrast to the other methods our approach does not use any validation set.

3.1. Unsupervised Google Re-Ranking Task

Previously, Sivic et al [24] used topic models on local feature representations for unsupervised learning. Fergus et al [8] extended their approach to encode spatial information. As the latter can be seen as the sparse counterpart to our dense representation, we compare on the unsupervised image re-ranking task specified in [8]. The provided data sets are results of image google queries. The task is to re-rank the images so that relevant ones appear first. The main challenge is to extract the relevant information which is hidden in an image set containing up to 70% junk images in an unsupervised fashion. Given that our representation effectively encodes the object structures, we expect our data to live in compact subspaces of the latent space. Therefore, we perform k-means clustering on the activations and consecutively accept the clusters with the most samples. The precision we obtain in this manner at 15% recall is shown in Table 1 and compared to our competitors. The average precision of 69% obtained by [8] and 54% obtained by [23] is surpassed by our approach which obtains an average precision of 91%. This performance is obtained without using the provided validation set which the other two approaches use. Although our method performs worst on the leopard data, we still improve over [8]. This is surprising as one would have suspected, that the local feature-based approach is more suited to encode the spotted texture of these animals. We account the success of our method to the added expressiveness by enabling the discovery of reoccurring contour fragments and edge segment like structures. Due to the dense and localized nature of our input features, we are more flexible to adapt to the object outline and to neglect background information. Figure 3 shows some topics from the presented experiment that expose these characteristics. Furthermore, in contrast to local feature-based methods our representation can easily be visualized (see Figure 3), which lend itself also to interaction and inspection by a user.

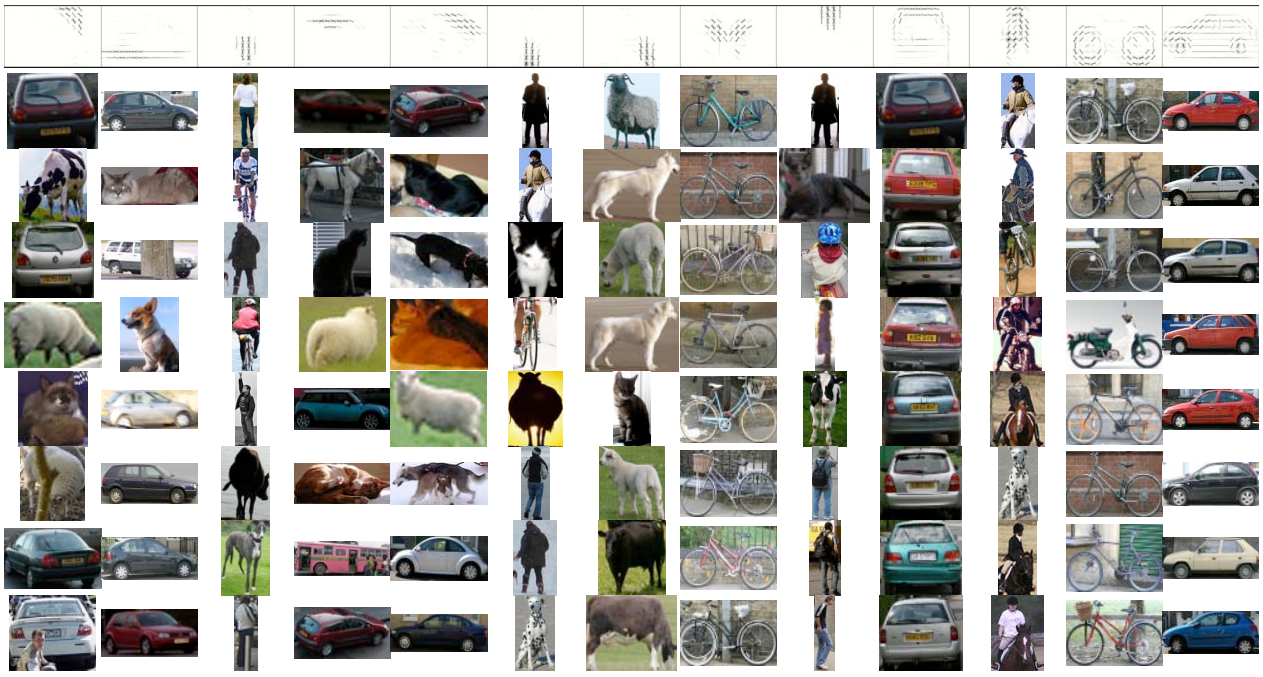


Figure 4. First row: example topics that were learned by the proposed approach across categories and viewpoints for the 10 classes of the PASCAL'06 data. Below first row: training images that activated the topic above most. The topics model local structures, line segments as well as silhouette-like structures. The topics are distinctive enough to separate several category members and even view-points. On the other hand they are general enough to be shared across categories and viewpoints.

3.2. Unsupervised Object Class Discovery

To extend our findings to the detection task that we are aiming for in Section 5.3, we extract our representation on the multi-category, multi-view PASCAL'06 dataset [7], in order to obtain a decomposition that is shared across categories.

In the first row of Figure 4 13 of 100 topic distributions are visualized that were trained on the bounding box annotations of the training and validation data of the PASCAL'06 challenge. The rows below display the examples that activated this particular topic most. We observe that the topics capture different levels of objects, ranging from global silhouettes (car rear in column 10 and side view in column 13) over localized parts (legs in column 3, bicycle frame in column 8 and bicycle wheels in column 12) to line segments and corners (corner in column 1 and line segments in column 2 and 4) . The model discovers distinctive parts that even separate several examples of different categories and their viewpoints although no such information was available to the system during training. Importantly, we can see that other topics like those that got activated on legs are shared across several categories, which is a desirable property of a compact decomposition in order to be scalable [27].

To illustrate that this is indeed an appropriate and effective approach to capture the variety of the data and to stress the power of modeling combinations of these discovered topics, we cluster the topic distributions as proposed in the last paragraph. Figure 5 shows in each row the 10 cluster members that are closest to the cluster center all of the 50 cluster centers. Keeping in mind that they are obtained in an entirely unsupervised fashion, the clusters turn out to be surprisingly clean.

We interpret these findings as strong evidence, that our model indeed captures an effective and low-dimensional representation for this difficult multi-category detection task.

4. Detection of Visual Categories

Based on the promising results on unsupervised learning in the last section, this section describes a complete system for supervised multi-category detection that leverages the learned representation.

4.1. Generative/Discriminative Training

Recently, the combinations of generative approaches with discriminative ones have shown to be very effective [13, 10]. The idea is that generative models can easily incorporate prior information to support learning from small samples, have increased robustness to noise and generally have more principled ways of dealing with missing data. Discriminative models on the other hand have shown to give superior performance for well posed learning tasks and

Figure 5. Unsupervised discovery of categories and viewpoints in PASCAL'06 data. The rows show for all 50 clusters those 10 examples that are closest to the cluster center.

a sufficient number of training examples. We also follow this idea and complement the generative model described in Section 2.2 by a discriminative SVM classifier with an RBF kernel [4]. In particular we train an SVM to discriminate between the topic distributions $\theta^{(d)}$ which are inferred for images containing the category of interest and others that do not contain these. By doing so, we seek to profit from the above mentioned benefits of the generative model combined with the discriminative classifier.

4.2. Sliding Window Approach to Detection

As proposed in [6] a sliding window approach can be done efficiently in this setting if the sliding window is always shifted by exactly one cell in x or y direction. In this case, the gradient histograms of the cell grid are computed once and for each sliding window the relevant sub grid is used.

Typically, sliding window techniques not only assign a high score for the correct location and scale in an image, but also for test windows that have a small offset in space and scale. We use a simple greedy scheme to cope with this issue: While there are unprocessed windows in an image, we accept the one with the highest score and reject all other windows that fulfill the symmetric overlap criterion

$$\max\left(\frac{A_i \cap A_j}{A_i}, \frac{A_i \cap A_j}{A_j}\right) > 0.3 \qquad (2)$$

where A_i and A_j are the areas covered by the two windows. As the bounding box scores from our approach turn out to be surprisingly consistent over different scales, this basic scheme has proven to work well in our setting.

Of course multi-scale detection task ranging over multiple octaves requires the investigation of large number of test windows – typically more than 10000 per image. While feature extraction and SVM classification are fast, our approach requires inference in the topic model for each test window rendering the method computationally infeasible for applications of interest. Therefore, we dedicate the following section to describe speed-ups that make our approach applicable to large databases.

4.3. Speed-ups: Linear Topic Response and Early Rejection

While we use the Gibbs sampling method [11] to estimate the model, we use the variational inference method described in [2] for test as it turns out to be computational more efficient given our setting. For more substantial improvements, we propose to compute a linear topic response to get an initial estimate on the topic activations. The aim is to avoid the more expensive inference scheme by performing an early rejection of the test windows. Different to linear methods like PCA, where there is linear dependency between the feature space and the coefficient space, the mixture coefficients of the topic distribution have to be fitted to the observation. This means that each word/feature can be associated to different topics depending on its context (presence of other features) and therefore also lead to strengthening or inhibition of other topic activations. This requires an iterative technique to find the best reconstruction. Therefore we ask the question how important this iterative fitting

and how much performance we loose by reverting to the following simple, linear approximation of the dependency between observed feature histogram x and topic activations $\theta^{(d)}$:

$$\tilde{\theta}^{(d)} = \left(\phi^{(1)} \ldots \phi^{(T)}\right)^t x, \quad (3)$$

In fact, our results on the UIUC single scale database show that there is a significant loss of about 8% in equal error rate performance (see Section 5.2), but a more detailed analysis on the UIUC multi-scale database reveals interesting results. Although, the linear approximation might be quite coarse, it can still be used for early rejection of test windows. It turns out, that full recall is achieved for the 2500 highest scored windows of a total of 2.826.783. As a consequence, more than 99.9% of the test windows can be handled by the linear computation that we measured to be 166 times faster than the proper inference. Taking all optimizations together we can cut down the computation time by a factor of 180 which corresponds to an reduction from one hour to around 20 seconds per image (AMD Opteron 270 (Dual-Core), 2.0 GHz).

5. Experiments

This section is divided into 4 parts. First, we show that our approach makes efficient use of the provided training examples by comparing to a baseline experiment on the UIUC single scale car database. Second, we evaluate different methods for estimation of the topic model on the UIUC multi-scale database and compare the obtained performance to previous work. Third, we present results on the PASCAL challenge 2006 data, that outperform the state-of-the-art on three of the ten categories. Fourth, we compare to a shape based approach on the ETH shape database to underline the versatility and adaptivity of our approach.

5.1. Efficient Use of Training Examples and Parameter Selection

To select parameters appropriate to our problem domain, we run detection experiments on the UIUC single scale car database which consists of a training set of 550 car and 500 background images of small size, while the test set has 170 images showing side views of cars in street scenes at a fixed scale. It turns out that the heuristic specified in [25] for selecting the hyperparameters α and β works very well for our setting. Therefore we use $\alpha = 50/\#topics$ and $\beta = 0.01$. We obtain best performance using 30 topics and a grid size of $(16, 6)$ for the gradient histograms.

To show that our approach makes efficient use of the provided training examples, we compare to a baseline experiment that does not use the proposed representation. Figures 6(a) and 6(b) show the precision-recall curves of our system, when trained on different numbers of positive and negative examples. We start with 50 car and 50 background images and increase by 50 until we use the full training dataset. The maximum performance is rapidly reached using only 150 positive and 150 negative examples. In contrast, the linear SVM trained on the same data representation but without our representation has a much slower learning curve. In fact the performance is 9.5% below the equal error rate of our new approach using 250 positive and 250 negative examples. We also tried RBF kernels, but obtained similar, inferior results.

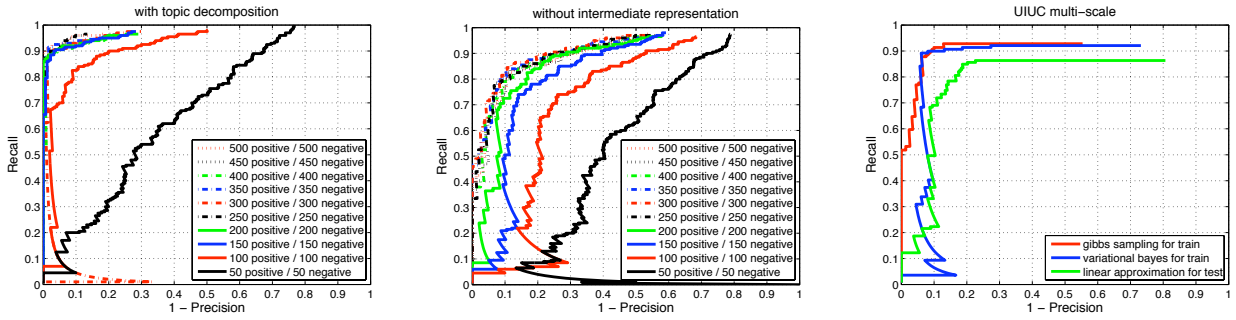
We account this significant improvement to the generative properties of our model inferring a generative decomposition of the presented data. We conclude, that this low dimensional representation simplifies the learning problem for the discriminative SVM classifier, which leads to more efficient use of training examples.

5.2. Comparison of Methods for Estimation and Evaluation of Approximate Inference

In this section we test the model that we trained for the UIUC single scale database on the multi-scale version and compare different estimation schemes for the topic model during training [2, 11]. We also evaluate the linear topic activations for testing that we proposed in Section 4.3. The results are reported in Figure 6(c). The estimation method based on Gibbs sampling [11] leads to similar performance as the variational inference method [2] when evaluated in the whole system, but shows better precision. We notice that the automatic selection of α that we use for the variational approach converged to a value of 0.373 which enforces less co-activation and therefore less sharing of topics. By visual inspection of the topic-distributions, we confirmed that the method of [2] learned more global topics, while the ones obtained by the Gibbs sampling method tends to be a little sparser. We believe that for detection tasks the second is to be preferred, as global representations can easier be mislead by effects like occlusion, as it is also supported by our results.

Replacing the proper inference by the linear approximation (Section 4.3) results in the third curve displayed in Figure 6(c). This confirms the importance and superiority of the proper inference in comparison to linear topic activations. For this comparison we use non-maxima suppression in combination with the linear approximation scheme while it is switched off when used for early rejection to achieve maximum recall.

The best results obtained by the Gibbs sampling approach with an equal error performance of 90.6% outperform [10] and are on par with the result in [20]. The best performance on this dataset have been reported by [28] with 93.5% and [19] with 94.7%, where the later used a different training set.



(a) (b) (c)

Figure 6. (a) and (b): Comparison of learning curve for proposed intermediate representation versus SVM on pure features on UIUC single-scale database. (c): Performance on UIUC multi-scale dataset using topic model estimated via Gibbs sampling vs variational bayes approach compared to using pseudo topic activations..

bicycle	bus	car	cat	cow	dog	horse	motorbike	person	sheep
49.75%	25.83%	50.07%	9.09%	15.63%	4.55%	9.40%	27.43%	0.98%	17.22%

Table 2. Average precision achieved on the PASCAL'06 database.

5.3. Comparison to state-of-the-art on PASCAL'06 VOC detection challenge

We evaluate our approach on the competition 3 of the PASCAL challenge 2006 [7] that poses a much harder detection problem as 10 visual categories are to be detected from multiple viewpoints over a large scale range.

We leave the hyperparameters untouched, but increase the number of topics to 100 and adapt the aspect ratio of the grid to $(16, 10)$. To reduce confusion between categories and the number of false positives, we adapt a bootstrapping strategy. First we train an initial model for each category versus the other categories. This model is then used to generate false positives on the training set (see also [21, 10, 6]). Up to 500 of the strongest false detection are added for each detector to its training set and the model is retrained. The average precisions of the final detector of all 10 categories on the test set are shown in Table 2 and the corresponding precision-recall curves are plotted in Figure 7. Figure 7 also shows some example detections of the system.

We outperform all other competitors in the 3 categories bicycle, bus and car by improving the state-of-the-art [7] on this dataset by 5.75%, 9.14% and 5.67% in average precision respectively. In particular we surpass the fully global approach [6] that our method was motived by. Compared to [5] we improve on bicycles and bus only by 0.65% and 0.93%, but again significantly on cars with 8.87%. However, in contrast to [5] we do not use the viewpoint annotations to train our approach. For the other categories, we perform about average, but also showed some inferior results on the highly articulated categories. We are currently investigating means to make the approach less rigid and carry over the good results from the first 3 categories to the other ones.

	Applelogos	Bottles	Giraffes	Mugs	Swans	average
our method	89.9%(4.5)	76.8%(6.1)	90.5 %(5.4)	82.7%(5.1)	84.0%(8.4)	84.8%
Ferrari [9]	83.2%(1.7)	83.2%(7.5)	58.6 %(14.6)	83.6 %(8.6)	75.4 %(13.4)	76.8%

Table 3. Comparison against shape-based approach of [9] on ETH shape database. Average detection-rate at 0.4 false positives per image averaged over 5-folds. Standard deviation is specified in brackets.

5.4. Comparison to shape features on ETH shape database

As pointed out in the previous experiments, our representation learns features with different characteristics from local to global and is in particular also capable of modeling contours. Therefore, we ask the question how our representation compares to shape-based approaches. We compare to [9] on the ETH shape database using the same detection system with the same settings as described in the last section. Example topics that were learnt across the 5 classes are depicted in Figure 3. Using five fold cross-validation as proposed in [9], we obtain the results presented in Table 3. Averaged over all classes we improve the performance of [9] by 8.0% to 84.8%. On applelogos, giraffes and swans, we improve the performance by 6.7%, 31.9% and 8.6% respectively. On mugs our approach performs comparable and on bottles it looses 6.4%. We account the worse performance on the bottles to the shape which is less discriminant with respect to the background. As the database was designed to test shape-based approaches, the improvements obtained by our approach underlines the versatility and adaptivity of the learnt representation.

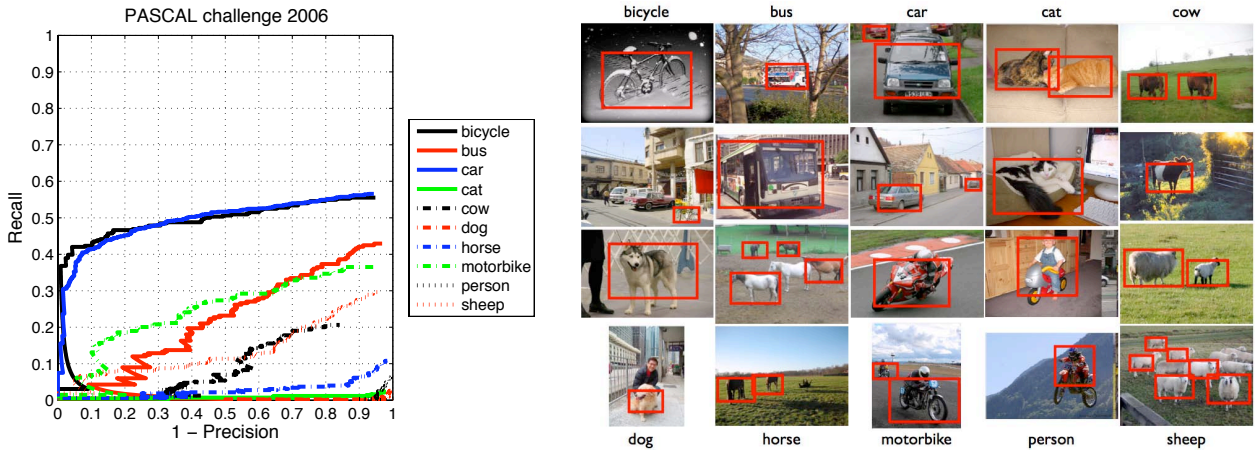


Figure 7. Results on the PASCAL VOC challenge 2006. Precision-Recall curves and example detections.

6. Conclusions

We present a novel method for representing multiple categories from multiple viewpoints and successfully employ it in various settings ranging from unsupervised learning to supervised detection tasks. In various experiments our approach shows superior performance with respect to purely local, shape-based or global approaches. Our representation has proven effective yet also efficient in showing an increased learning curve in the detection setting. Beyond the modeling aspects, we pay particular attention to computational feasibility that enables scalability to large databases. Lastly, we want to highlight the results on the challenging PASCAL'06 dataset where we improve the state-of-the-art on three categories to underline our contribution to category modeling in the context of a complete detection system.

Acknowledgments: We thank Rob Fergus for helpful discussions. We gratefully acknowledge support by the Frankfurt Center for Scientific Computing. This work has been funded, in part, by the EU project CoSy (IST-2002-004250).

References

[1] A. Bissacco, M.-H. Yang, and S. Soatto. Detecting humans via their pose. In *NIPS*, 2007.

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

[3] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *ECCV*, 2006.

[4] C.-C. Chang and C.-J. Lin. *LIBSVM: a library for support vector machines*, 2001. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[5] O. Chum and A. Zisserman. An exemplar model for learning object classes. In *CVPR*, 2007.

[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *ICCV*, 2005.

[7] M. Everingham, A. Zisserman, C. K. I. Williams, and L. Van Gool. The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. `http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf`.

[8] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google"s image search. In *ICCV*, 2005.

[9] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid. Accurate object detection with deformable shape models learnt from images. In *CVPR*, 2007.

[10] M. Fritz, B. Leibe, B. Caputo, and B. Schiele. Integrating representative and discriminant models for object category detection. In *ICCV*, 2005.

[11] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS USA*, 2004.

[12] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001.

[13] T. S. Jaakkola and D. Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, 1998.

[14] I. Laptev. Improvements of object detection using boosted histograms. In *BMVC*, 2006.

[15] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization. In *BMVC*, 2006.

[16] B. Leibe, E. Seemann, and B. Schiele. Pedestrian detection in crowded scenes. In *CVPR*, 2005.

[17] F.-F. Li and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, 2003.

[19] K. Mikolajczyk, B. Leibe, and B. Schiele. Multiple object class detection with a generative model. In *CVPR*, 2006.

[20] J. Mutch and D. G. Lowe. Multiclass object recognition with sparse, localized features. In *CVPR*, 2006.

[21] E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *CVPR*, 1997.

[22] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *ICCV*, 2005.

[23] F. Schroff, A. Criminisi, and A. Zisserman. Harvesting image databases from the web. In *ICCV*, 2007.

[24] J. Sivic, B. Russell, A. A. Efros, A. Zisserman, and B. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[25] M. Steyvers and T. L. Griffiths. Probabilistic topic models. In *Handbook of Latent Semantic Analysis*. Lawrence Erlbaum Associates, 2007.

[26] E. Sudderth, A. Torralba, W. Freeman, and A. Willsky. Learning hierarchical models of scenes, objects, and parts. In *ICCV*, 2005.

[27] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *PAMI*, 29(5), 2007.

[28] B. Wu and R. Nevatia. Cluster boosted tree classifier for multi-view, multi-pose object detection. In *ICCV*, 2007.