Constructing visual models with a latent space approach

Florent Monay, Pedro Quelhas, Daniel Gatica-Perez, and Jean-Marc Odobez

monay@idiap.ch,quelhas@idiap.ch,gatica@idiap.ch,odobez@idiap.ch
IDIAP Research Institute, 1920 Martigny, Switzerland

Abstract. We propose the use of latent space models applied to local invariant features for object classification. We investigate whether using latent space models enables to learn patterns of visual co-occurrence and if the learned visual models improve performance when less labeled data are available. We present and discuss results that support these hypotheses. Probabilistic Latent Semantic Analysis (PLSA) automatically identifies aspects from the data with semantic meaning, producing unsupervised soft clustering. The resulting compact representation retains sufficient discriminative information for accurate object classification, and improves the classification accuracy through the use of unlabeled data when less labeled training data are available. We perform experiments on a 7-class object database containing 1776 images.

1 Introduction

The bag-of-words model is one of the most common text document representations in information retrieval (IR), in which a fixed-size vector stores the occurrence of the words present in a document. Although the sequential relations between words are not preserved, this somewhat drastic simplification allows simple comparisons between documents, and produces good performance for document classification and retrieval [1].

A text corpus represented by a bag-of-words is an example of a collection of discrete data, for which a number of generative probabilistic models have been recently proposed [5, 2, 3, 6]. The models, able to capture co-occurrence information between word and documents, have shown promising results in text dimensionality reduction, feature extraction, and multi-faceted clustering. It is thus not a surprise that the interest in casting other data sources into this representation has increased; recent work in computer vision has shown that images and videos are suitable for a vector-space representation, both for visual tasks like object matching [14], object classification [17], and cross-media tasks like image auto-annotation [4, 9, 10].

We propose here to build visual models from images in a similar fashion, using a quantized version of local image descriptors, dubbed visterms [15, 14]. However, unlike related work, which has only used the basic bag-of-words [14, 17], we propose to use a probabilistic latent space model, namely Probabilistic Latent Semantic Analysis (PLSA) [5] to build visual models of objects.

The different outcomes of this model are principally unsupervised feature extraction and automatic soft clustering of image datasets, that we recently studied in the context of *scene modeling* [12]. Independently, Sivic et al. compared two latent probabilistic models of discretized local descriptors to discover object categories in image collections [13]. The approach is closely related to what we propose in this paper and in [12], but fundamentally differs in the assumption of the latent structure of the data. In [13], the number of classes is assumed to be known a priori. In contrast we assume that an image is a mixture of latent aspects that are not necessarily limited to the number of object categories in the dataset. We consider latent aspect modeling not as a classification system in itself, but as a feature extraction process for supervised classification. We show (qualitatively and quantitatively) the benefits of our formulation, and its advantages over the simple vector-space formulation. Based on the results, we believe that the approach might be worth exploring in other vision areas.

The paper is organized as follows. Section 2 describes the specific probabilistic model. In Section 3 we discuss the image representation. Section 4 summarizes results regarding object clustering and classification, and Section 5 concludes the discussion.

2 Latent structure analysis

2.1 Bag-of-words: data sparseness

The vector-space approach tends to produce high-dimensional sparse representations. Sparsity makes the match between similar documents difficult, especially if ambiguities exist in the vector-space. In the text case for example, different words might mean the same (synonymy) and a word can have several meanings (polysemy). This potentially leads to ambiguous data representations. In practice, such situation also occurs with visterms.

To overcome this problem, different probabilistic generative models [5, 2, 3, 6] have been proposed to learn the co-occurrence between elements in the vector-space in an unsupervised manner. The idea is to model a latent data structure from the co-occurrence of elements in a specific dataset, assuming their independence given a latent variable. The elements in the vector space are probabilistically linked through the latent *aspect* variable, which identifies a disambiguated lower-dimensional representation. One model that implements this concept is PLSA, which we briefly review in the following.

2.2 Probabilistic LSA

In a dataset of N_d documents represented as bag-of-words of size N_x, the PLSA model assumes that the joint probability of a document d_i and an element x_j from the vector-space is the marginalization of the N_z joint probabilities of d_i,

x_j and an unobserved latent variable z_k called *aspect*:

$$P(x_j, d_i) = \sum_{k=1}^{N_z} P(x_j, z_k, d_i)$$
$$= P(d_i) \sum_{k=1}^{N_z} P(z_k \mid d_i) P(x_j \mid z_k). \tag{1}$$

Each document is a mixture of latent aspects, expressed by the conditional probability distribution of the latent aspects given each document d_i, $P(z \mid d_i)$. Each latent aspect z_k is defined by the conditional probability distribution $P(x \mid z_k)$ in Eq. 1. The parameters are estimated by the Expectation-Maximization (EM) procedure described in [5] which maximizes the likelihood of the observation pairs (x_j, d_i). The E-step estimates the probability of the aspect z_k given the element x_j in the document d_i (Eq. 2).

$$P(z_k \mid d_i, x_j) = \frac{P(x_j \mid z_k) P(z_k \mid d_i)}{\sum_{k=1}^{N_z} P(x_j \mid z_k) P(z_k \mid d_i)} \tag{2}$$

The M-step then derives the conditional probabilities $P(x \mid z_k)$ (Eq. 3) and $P(z \mid d_i)$ (Eq. 4) from the estimated conditional probabilities of aspects $P(z_k \mid d_i, x_j)$ and the frequency count of the element x_j in image d_i, $n(d_i, x_j)$.

$$P(x_j \mid z_k) = \frac{\sum_{i=1}^{N_d} n(d_i, x_j) P(z_k \mid d_i, x_j)}{\sum_{m=1}^{N_x} \sum_{i=1}^{N_d} n(d_i, x_m) P(z_k \mid d_i, x_m)} \tag{3}$$

$$P(z_k \mid d_i) = \frac{\sum_{j=1}^{N_x} n(d_i, x_j) P(z_k \mid d_i, x_j)}{n(d_i)} \tag{4}$$

To prevent over-fitting, the number of EM iterations is controlled by an early stopping criterion based on the validation data likelihood. Starting from a random initialization of the model parameters, the EM iterations are stopped when the criterion is reached. The corresponding latent aspect structure defined by the current conditional probability distributions $P(x \mid z_k)$ is saved. Derived from the vector-space representation, the inference of $P(z_k \mid d_i)$ can be seen as a feature extraction process and used for classification. It also allows to rank images with respect to a given latent aspect z_k, which illustrates the latent structure learned from the data.

3 Images as bag-of-visterms

Although global features such as global color histograms or global edge direction histograms are traditionally used to represent images, a promising recent research direction in computer vision is the use of local image descriptors. The

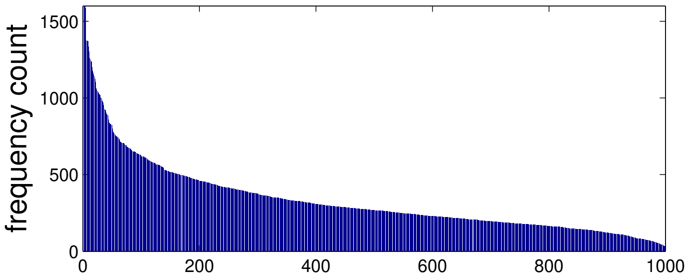

Fig. 1. Sorted document frequency counts of the quantized local image patches in the training set.

combination of interest point detectors and invariant local descriptors has shown interesting capabilities of describing images and objects. We decided to use the Difference of Gaussians (DOG) point detector [7] and the Scale Invariant Feature Transform (SIFT) local descriptors [7], as proposed in recent studies [8]. The SIFT descriptors are local histograms of edge directions and therefore correspond to local image structures. Note that only gray-level information is used for this process.

The idea is to identify different types of local image patches occurring in the database to represent an image, similarly to the bag-of-words approach. As for the word ordering, the spatial information of the local descriptors is not encoded in the image representation. Those local image patches are obtained by a standard K-means quantization of the extracted SIFT descriptors in an image dataset, and are referred to as *visterms* (visual terms). As an analogy with text, the image representation is referred to as *bag-of-visterms* (BOV). We did not experiment the standard inverse document frequency (idf) weighting, but restricted our experiments to the unweighted BOV representation. As shown in Figure 3, the K-means quantization produces much more balanced document frequencies than what is encountered in text (Zipf's law), and the BOV representation therefore does not need to be compensated.

4 Image modeling with PLSA

4.1 Data description

To create the visterm vocabulary (K-means) we use a 3805-image dataset constructed from several sources. This includes 1002 building images (Zubud), 144 images of people and outdoors [11], 435 indoor images with people faces [17], 490 indoor images from the corel collection [16], 1516 city-landscape overlapped images from Corel [16] and 267 Internet photographic images. Interests points are identified on each image with the DOG point detector, a SIFT description of each point is computed and all SIFT descriptors are quantized with K-means to construct the visterms 'vocabulary'.

We propose to consider a 7-class dataset to evaluate classification [17]. The image classes are: faces (792), buildings (150), trees (150), cars (201), phones (216), bikes (125) and books (142), adding up to a total of 1776 images. The size of the images varies considerably: images can have between 10k and 1,2M pixels while most image sizes are around 100-150k pixels. We resize all images to 100k pixels since the local invariant feature extraction process is highly dependent of the image size. This ensures that no class-dependent image size information is included in the representation. The dataset is split in 10 test sets, which allows ten evaluation runs with different training and test sets each time. We decided to use 1000 visterms to represent each image (size of the BOV).

4.2 Image soft clustering

The latent structure learned by PLSA can be illustrated by the top-ranked images in a dataset with respect to the posterior probabilities $P(z_k \mid d_i)$. Fig. 2 shows a ranking of seven out of 20 aspects identified by PLSA on the 7-class dataset described above. We selected N_z= 20 for a cleaner ranking visualization. From Fig. 2, we observe that aspects 3 and 17 seem closely related to face images. The first ten images ranked with respect to aspect 8 are all bike images, while top-ranked images for aspect 10 mostly contain phones. Buildings are present in aspect 5, all images related to aspect 7 are tree images. Aspect 12 does not seem to be related to any specific object category.

To analyze the ranking in more details, the precision and recall curves for the retrieval of faces, cars, bikes, and trees are shown in Fig. 3. The top left graph shows that the homogeneous ranking holds on for more than 10 retrieved images in aspect 3 and 17, confirming the observations made from Fig. 2. We see that another aspect (13) is closely related to face images. The top right graph from Fig. 3 shows that aspect number 12 is related to car images if looking deeper in the ranking, what is not obvious from the observation of Fig. 2. Note however that the precision/recall values are not as high as for the faces case. The bottom left graph confirms that aspect 8 is linked to bike images, as well as aspect 1 even if less obvious. The bottom right graph shows that top-ranked images with respect to aspect 7 are mainly tree images. These results confirm that PLSA can capture class-related information in an unsupervised manner.

4.3 Images as mixtures of aspects

Our model explicitly considers an image as a mixture of latent aspects expressed by the $P(z \mid d)$ distributions learned from PLSA. The same latent structure with N_z= 20 aspects used for the aspect-based image ranking is considered. As illustrated by the aspect-based image ranking from Fig. 2, some identified aspects relate to specific object categories. Within the dataset, different examples of aspect mixtures can be observed. In Fig. 4 (a) the aspect distribution is mainly concentrated on the aspect related to 'building' images. The image only contains

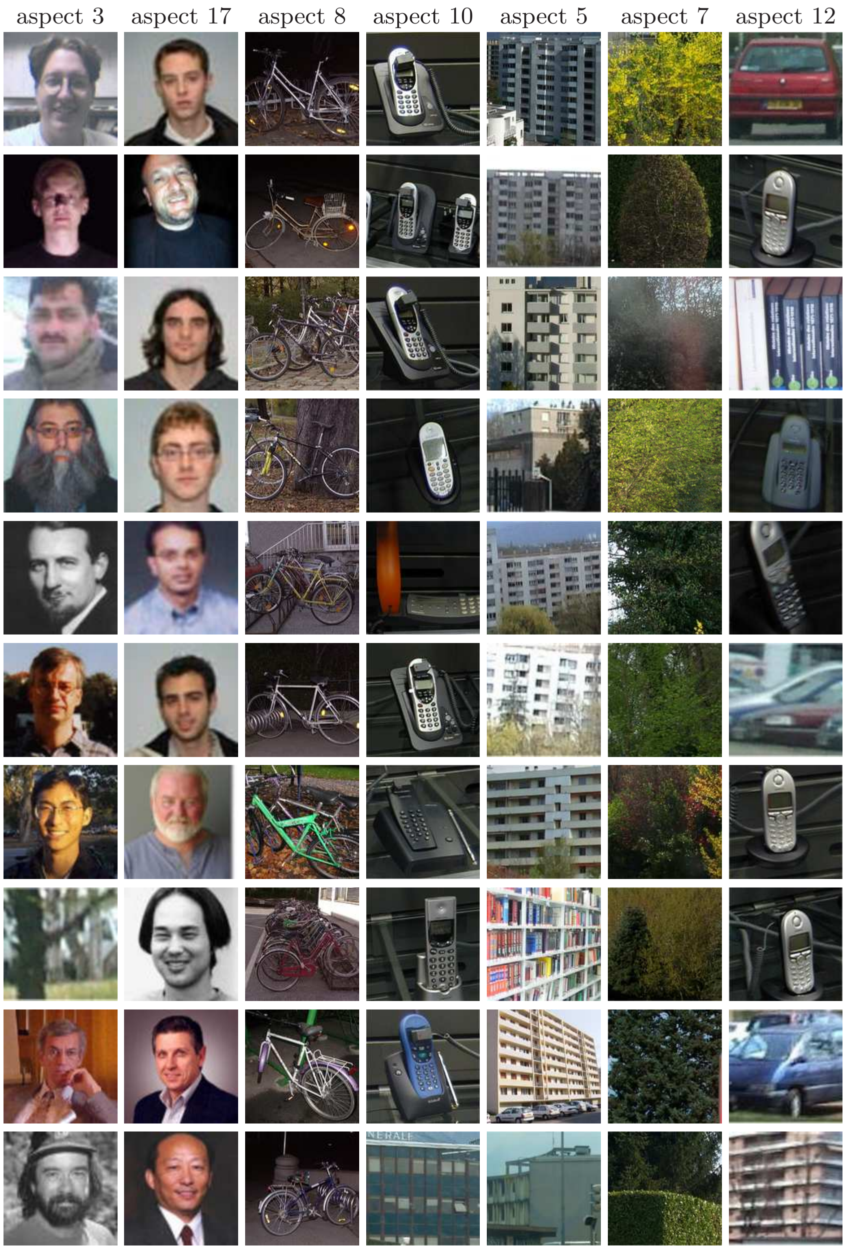

Fig. 2. 10 top-ranked images with respect to $P(z_k \mid d_i)$ for seven selected aspects. Images are cropped for a convenient display. A full ranking is available at http://www.idiap.ch/∼monay/PASCAL_LATENT/

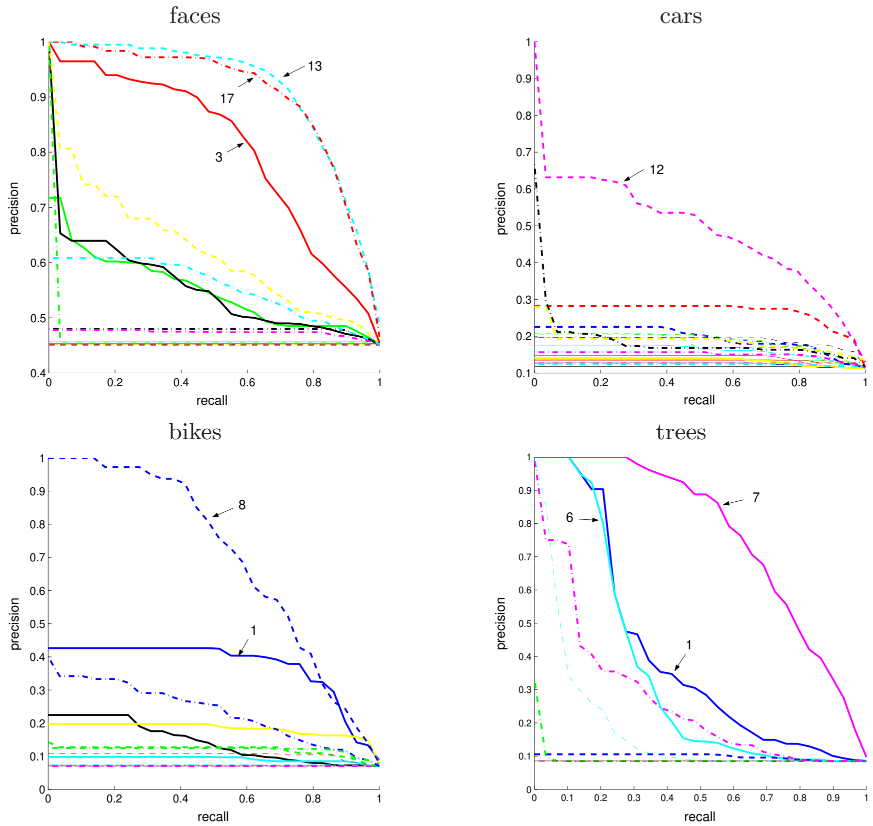

Fig. 3. Precision and recall curves for the 'face', 'car', 'bike' and 'tree' categories, according to an aspect-based unsupervised image ranking. The lowest precision values on the graph correspond to a random ranking.

building structures, therefore the aspect distribution seems coherent. On the contrary, the image from Fig. 4 (b) is composed of both 'building' and 'tree' -related structures. The corresponding aspect distribution interestingly reflects this image composition with the most probable aspects related to 'building' and 'tree'.

It is important to point out that there are cases when the aspect distribution does not clearly correspond to the image semantic. Fig. 5 (a) shows the close-up of a bike, but the aspect distribution is not concentrated on aspect 8, previously related to 'bike' images. The aspect distribution $P(z \mid d)$ rather describes the image as a mixture of several aspects with no specific dominance. This ambiguous aspect representation could derive from the fact that only a few examples of this type of close-up appear in the database. In Fig. 5 (b), the image is identified

(a)

(b)

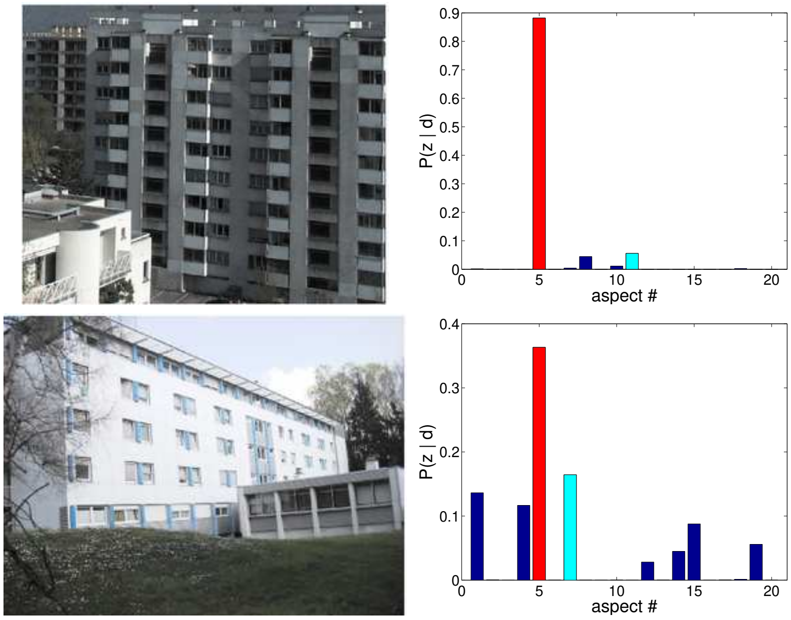

Fig. 4. Images and their corresponding aspect distribution $P(z \mid d)$ for N_z= 20. (a) is concentrated on aspect 5 (building), while (b) is a mixture of aspects 5 (building), 7 (tree) and aspect 1.

as a mixture of aspect 8 and 7, which perfectly reflects the image composition. Bikes are located in the image on a tree/vegetation background.

4.4 Feature extraction

The PLSA model can be seen as a feature extraction or dimensionality reduction process: from the bag-of-visterms, a lower-dimensional aspect-based representation $P(z_k \mid d_i)$ is inferred using a previously learned PLSA model. Here we propose to compare the aspect-based and the bag-of-visterms representations on the 7-class supervised classification task. The PLSA model is trained on all non-test images each time and the resulting model is used to extract the aspect-based representation. To evaluate the quality of the feature extraction, we compare the classification based on the BOV representation with the aspect-based representation with the same classification setup: one Support Vector Machine (SVM) per class is trained with one class against all others.

Table 1 and Table 2 show the confusion matrix for the BOV and the PLSA-based classification with N_z= 60 aspects. The last column is the per class error. We see that the classification performance greatly depends on the object class for both the BOV and the PLSA representations. These differences are caused by diverse factors. For instance 'trees' is a well defined class that is dominated

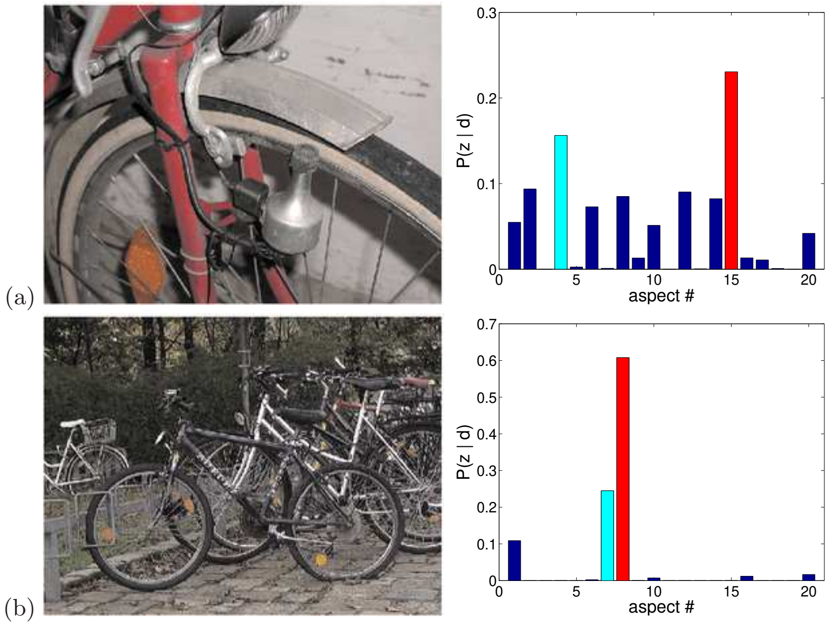

Fig. 5. Images and their corresponding aspect distribution $P(z \mid d)$ for N_z= 20. (a) is a mixture of different aspects, (b) is a mixture of aspect 8 (bikes) and 7 (trees).

by high frequency texture visterms, and therefore does not get confused with other classes. Similarly, most 'face' images have an homogeneous background and consistent layout which will not create ambiguities with other classes in the BOV representation. This explains the good performance of these two categories.

On the contrary, 'car' images present a large variability in appearance within the database. Front, side and rear car views on different types of background can be found, what makes it a highly complex category for object classification, generating an important confusion with other classes. 'Phones', 'books' and 'buildings' are therefore confused with 'cars' in both the BOV and the PLSA case. The 'bike' class is well classified despite a variability in appearance comparable to the 'car' images, because the bike structure generates a discriminative BOV representation.

Table 3 summarizes the whole set of experiments when we gradually train the SVM classifiers with less training data. If using all the training data (90% of all data) for feature extraction and classification, BOV and PLSA achieve a similar total error score. This proves that while achieving a dimensionality reduction from 1000 visterms to N_z= 60 aspects, PLSA keeps sufficient discriminative information for the classification task.

The case in which PLSA is trained on all the training data, while the SVMs are trained on a reduced data portion of it, it corresponds to a partially labeled data problem. Being completely unsupervised, the PLSA approach can take

	faces	buildings	trees	phones	cars	bikes	books	error
faces	772	2	7	3	3	2	3	2.5(0.04)
buildings	6	100	6	5	12	5	16	33.3(1.70)
trees	1	3	141	1	3	1	0	6.0(0.60)
phones	14	0	0	187	6	2	7	13.4(1.20)
cars	18	1	2	12	162	3	3	19.4(1.46)
bikes	0	3	3	1	2	116	0	7.2(0.38)
books	13	8	0	9	9	1	102	28.2(1.86)

Table 1. Confusion matrix for the 7-class object classification problem using the bag-of-visterms features, summed over 10 runs, and average classification error with the variance over ten runs indicated in brackets.

	faces	buildings	trees	phones	cars	bikes	books	error
faces	772	2	5	1	10	1	1	2.5(0.02)
buildings	2	113	3	3	18	5	6	24.6(1.40)
trees	3	3	140	0	2	2	0	6.7(0.40)
phones	9	5	0	166	23	2	11	23.1(0.60)
cars	14	5	0	3	172	4	3	14.4(0.67)
bikes	0	3	4	0	4	113	1	9.6(0.69)
books	7	13	0	6	14	0	102	28.2(1.54)

Table 2. Confusion matrix for the 7-class object classification problem using PLSA with N_z= 60 aspects as a feature extraction process, summed over 10 runs, and average classification error with the variance over ten runs indicated in brackets.

advantage of any unlabeled data and build the aspect-based representation from it. This advantage with respect to supervised strategies is shown in Table 3 for 50%, 10% and 5% training data. Here the comparison between BOV and PLSA is done for the same reduced number of labeled images to train the SVM classifiers, while the PLSA model is still trained on the full 90% training data. The total classification errors show that the features extracted by PLSA outperform the raw BOV representations for the same amount of labeled data. Note also that the variance over the splits is smaller, which suggests that the model is more stable given the reduced dimensionality.

Method	90%	50%	10%	5%
PLSA (N_z= 60)	11.1(1.6)	12.5(1.5)	18.1(2.7)	21.7(1.7)
BOV	11.1(2.0)	13.5(2.0)	21.8(3.6)	26.7(2.8)

Table 3. Comparison between the bag-of-visterms (BOV) and the PLSA-based representation (PLSA) for classification with an SVM classifier trained with progressively less training data on the 7-class problem. The number in brackets is the variance over the different data splits.

5 Conclusion

For an object classification task, we showed that using PLSA on a bag-of-visterms representation (BOV) produces a compact, discriminative representation of the data, outperforming the standard BOV approach in the case of small amount of training data. Also, we showed that PLSA can capture semantic meaning in the BOV representation allowing for both unsupervised ranking of object images and description of images as a mixture of aspects. These results motivate further investigation of this and other latent space approaches for task related to object recognition.

6 Acknowledgments

This work was also funded by the European project "CARTER: Classification of visual Scenes using Affine invariant Regions and TExt Retrieval methods" part of "PASCAL: Pattern Analysis, Statistical Modeling and Computational Learning", through the Swiss Federal Office for Education and Science (OFES).

We thank the Xerox Research Center Europe (XRCE) and the University of Graz for collecting the object images and making the database available in the context of the Learning for Adaptable Visual Assistant (LAVA) project.

The authors acknowledge financial support provided by the Swiss National Center of Competence in Research (NCCR) on Interactive Multimodal Information Management (IM)2. The NCCR is managed by the Swiss National Science Foundation on behalf of the Federal Authorities.

References

1. R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* ACM Press, 1999.
2. D. Blei, Y. Andrew, and M. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1020, 2003.
3. W Buntine. Variational extensions to em and multinomial pca. In *Proc. of Europ. Conf. on Machine Learning*, Helsinki, Aug. 2002.
4. P. Duygulu, K. Barnard, N. Freitas, and D. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proc. of IEEE Europ. Conf. on Computer Vision*, Copenhagen, Jun. 2002.
5. T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42:177–196, 2001.
6. M. Keller and S. Bengio. Theme topic mixture model: A graphical model for document representation. *IDIAP Research Report, IDIAP-RR-04-05*, January 2004.
7. D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60:91–110, 2003.
8. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Madison, Jun. 2003.
9. F. Monay and D. Gatica-Perez. On image auto-annotation with latent space models. In *Proc. of ACM Int. Conf. on Multimedia*, Berkeley, Nov. 2003.

10. F. Monay and D. Gatica-Perez. PLSA-based image auto-annotation: Constraining the latent space. In *Proc. ACM Int. Conf. on Multimedia*, New York, Oct. 2004.
11. A. Opelt, M. Fussenegger, A. Pinz, and P. Auer. Weak hypotheses and boosting for generic object detection and recognition. In *Proc. of IEEE Europ. Conf. on Computer Vision*, Prague, May 2004.
12. 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 *Proc. of IEEE Int. Conf. on Computer Vision*, Beijing, Oct. 2005.
13. J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering object categories in image collections. Technical report, Dept. of Engineering Science, University of Oxford, 2005.
14. J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *Proc. of IEEE Int. Conf. on Computer Vision*, Nice, Oct. 2003.
15. T. Tuytelaars and L. Van Gool. Content-based image retrieval based on local affinely invariant regions. In *Proc. of Visual99*, Amsterdam, Jun. 1999.
16. A. Vailaya, M. Figueiredo, A. Jain, and H.J. Zhang. Image classification for content-based indexing. *IEEE Trans. on Image Processing*, 10:117–130, 2001.
17. J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *Proc. of ICPR Workshop on Learning for Adaptable Visal Systems*, Cambridge, Aug. 2004.