Universität Augsburg

Towards Universal Visual Vocabularies

C. Ries, S. Romberg, R. Lienhart

Report 2010-05 Juli 2010

Institut für Informatik
D-86135 Augsburg

TOWARDS UNIVERSAL VISUAL VOCABULARIES

Christian X. Ries, Stefan Romberg, Rainer Lienhart*

University of Augsburg
Email: {ries,romberg,lienhart}@informatik.uni-augsburg.de

ABSTRACT

Many content-based image mining systems extract local features from images to obtain an image description based on discrete feature occurrences. Such applications require a visual vocabulary also known as visual codebook or visual dictionary to discretize the extracted high-dimensional features to visual words in an efficient yet accurate way.

Once such an application operates on images of a very specific domain the question arises if a vocabulary built from those domain-specific images needs to be used or if a ”universal” visual vocabulary can be used instead. A universal visual vocabulary may be computed from images of a different domain once and then be re-used for various applications and other domains.

We therefore evaluate several visual vocabularies from different image domains by determining their performance at pLSA-based image classification on several datasets. We empirically conclude that vocabularies suit our classification tasks equally well disregarding the image domain they were derived from.

Keywords— visual vocabulary, visual words, image classification

1. INTRODUCTION

In recent years bag-of-words models have become increasingly popular for image content analysis due to the power to provide a useful image description and yet handle occlusion and some minor transformations effectively. Other techniques like computing topic models such as probabilistic latent semantic analysis (pLSA) on top of a bag-of-words model further improve the expressiveness of the image descriptions, and are widely used for many applications such as image retrieval and image classification.

During the development of a reliable adult image filtering system in our lab an interesting question has been raised: Does one have to create a visual vocabulary that is built from images of the target domain or can a vocabulary computed from generic images be used instead?

In this paper we therefore empirically examine whether bag-of-words models always require the time-consuming computation of distinct visual vocabularies. For this purpose, we analyze the classification results of such models based on vocabularies from different image domains.

Our experiments suggest that the expense of creating a new visual vocabulary over and over again can be omitted since one can re-use a universal visual vocabulary. We experimentally verify this statement for different vocabulary sizes and feature types.

2. RELATED WORK

To create our visual vocabularies, we cluster image features by employing hierarchical vocabulary trees as introduced by Nister et al. [1]. The pLSA topic model we use to derive a high-level image description is an unsupervised approach to compute topic distributions of documents from simple word occurrences. It was introduced by Hofmann [2][3] in the realm of text analysis.

Bosch et al. [4] proposed an approach to scene classification based on pLSA models.

Lienhart et al. [5] applied the pLSA to the problem of image retrieval in large scale image databases in a similar way and showed that using topic vectors to represent images yields a more accurate image description than plain feature histograms.

The image classification system we use for experimental evalutation of our visual vocabularies is based on the work of Hörster et al. [6] where the high-level image descriptions computed by the pLSA model are used to classify images.

3. CONTRIBUTIONS

In this paper we show empirically that the performance of bag-of-words models does not depend on the set of images used for the creation of the underlying visual vocabulary (as long as we use a reasonable number of images and a diverse enough image set from which we derive the vocabulary).

We also examine if this conclusion holds for vocabularies containing different numbers of words. Therefore, we create each visual vocabulary in three different sizes, that is we compute vocabularies consisting of 500, more than 9000 and more than 90,000 visual words for each of four image sets. We use pLSA image classification to evaluate the vocabularies by comparing the classification performance for each vocabulary.

To ensure that the observations are independent from the visual feature type used for creating the visual vocabularies, we

*This work was kindly funded by **Advanced U.S. Technology Group** (ATG), http://www.austg.com. Special thanks to CEO Dr. Rudolf Hauke.

compare Self-Similarity features [7] to SIFT features [8] by creating each of our vocabularies for both feature types.

Our results suggest that there is no need to compute a new visual vocabulary each time we want to create a bag-of-words model for a specific image domain. In fact we can simply compute one general visual dictionary and can re-use it for all image domains. Therefore the computation time for such models can be reduced significantly.

4. VISUAL VOCABULARIES

This section describes the four image sets we use to compute visual vocabularies. We also briefly explain the used features and the vocabulary creation.

4.1. Image Sets

We use four different image collections to compute visual vocabularies in several configurations and to evaluate their impact on the performance of the actual classification task.

- *Real world images: Flickr*
  The first image set consists of 100,000 images downloaded from the Flickr community image database by querying the Flickr web service using more than 25 different query terms (e.g. 'baseball', 'flowers', ...). Statistical analysis shows that about 40% of all Flickr images labeled with a certain tag do not actually show the expected content. Therefore, this image set can be considered a heterogeneous collection of real world images with both diverse content and appearance.

- *Artificial object categories: Caltech 101*
  Our second image set is the complete Caltech 101 dataset [9] which consists of about 9000 images from 101 object classes and one background class. The number of images per category varies from 40 to 800, but most categories consist of about 50 images. This dataset features a high number of noise-free image classes. Caltech 101 is commonly not considered a challenging dataset.

- *Domain-specific images: adult images*
  Since the original question was if it is necessary to compute domain-specific vocabularies for the application of adult image filtering, we use a collection of 2600 adult images to examine the application-specific impact.

- *Domain-specific images: bikini images*
  We further evaluate the image classification and the vocabularies on a collection of 2600 slightly noisy bikini images (i.e. images featuring women wearing bikinis) as they are somewhat inbetween inoffensive images and adult images. Again, this tests the actual impact of the vocabulary on a domain-specific classification task.

  The domain of bikini images is necessary as mostly one cannot work with the real adult images due to legal reasons, and the classification system has to be developed without actually ”knowing” the target domain. It is further of interest if vocabularies from a domain closely related to the adult images perform better at classifying these images than vocabularies from unrelated domains like the Flickr images.

Table 1. The image sets used to compute the visual vocabularies and the respective numbers of images

Vocabularies with 500 visual words	
Flickr:	1000 random Flickr images
Flickr/Bikini:	500 Flickr images, 500 bikini images
Flickr/Adult:	500 Flickr images, 500 adult images
Caltech 101:	4500 images (50% of each class)

Vocabularies with 9,000+ visual words	
Flickr:	1000 random Flickr images
Flickr/Bikini:	500 Flickr images, 500 bikini images
Flickr/Adult:	500 Flickr images, 500 adult images
Caltech 101:	4500 images (50% of each class)

Vocabularies with 90,000+ visual words	
Flickr	5000 random Flickr images
Flickr/Bikini:	2500 Flickr images, 2500 bikini images
Flickr/Adult:	2500 Flickr images, 2500 adult images
Caltech 101:	4500 images (50% of each class)

In order to obtain a reasonable number of features to cluster, we randomly draw subsets for which we then extract local features and build feature histograms for further processing. These subsets are listed in Table 1.

4.2. Local Image Features

We use two different types of local feature descriptors: SIFT and Self-Similarity.

The *Self-Similarity* descriptor was introduced by Shechtman et al. [7]. It models the similarity of a small center image patch to the image patches in its surrounding. These similarities are represented as correlation surfaces which are then transformed to a log-polar representation.

The *SIFT* features which were developed by Lowe [8] are among the best-known descriptors. SIFT features are basically weighted gradient histograms. Originally they were designed for use with the Difference-of-Gaussian keypoint dectector to find stable keypoints in order to recognize and match objects across images. However we extract SIFT features at locations given by a dense grid as recent work [6] showed that this yields better performance.

We compute both Self-Similarity and SIFT features on a dense grid of 5×5 pixels from images with a maximum side length of 640 pixels, i.e. larger images are scaled down. By evaluating two different feature descriptors we show empirically that our results do not depend on a certain feature type.

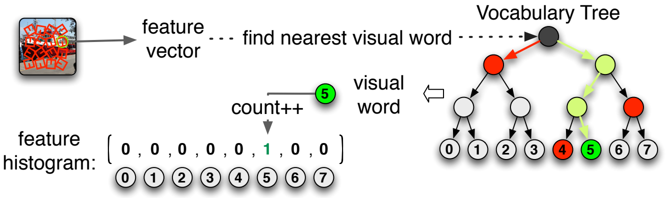

Fig. 1. Usage of a vocabulary tree to quantize feature vectors into discrete visual words in order to construct the document vector (word occurrence histogram).

4.3. Clustering

Most applications need to derive a discrete set of visual words from the actual extracted visual features. One possible approach is to partition a set of feature vectors into homogeneous groups or clusters. A representative feature such as the centroid is then chosen for each cluster and called a visual word. The set of representative feature centroids then defines the visual vocabulary. An unknown feature vector can then be mapped to its nearest representative feature, which allows computing the discrete visual word occurrences.

One common approach to compute the visual vocabulary is to cluster the feature vectors with the k-means algorithm [10]. This algorithm is a widely-used clustering algorithm and allows unsupervised clustering of multidimensional data into k groups. It clusters vectors iteratively into groups by minimizing the distance of each sample vector towards its nearest cluster centroid. While the k-means algorithm is quite fast for a small number of clusters k, it is time-consuming for larger vocabularies. In our experiments we use it to build the vocabularies that consist of 500 visual words.

For vocabularies of larger sizes using a flat vocabulary is expensive both at creation and at quantizing features efficiently. Therefore, to examine our larger vocabularies with more than 9,000 and 90,000 words we use hierarchical k-means clustering to cluster the extracted features.

Hierarchical k-means clustering produces a vocabulary tree [1] that hierarchically partitions the feature space. The creation of this tree is done by applying the k-means algorithm repeatedly to partition the feature space recursively into a small number of subspaces on each tree level.

Vocabulary trees have two advantages: They can be computed very fast as on every tree level the k-means algorithm only has to compute a coarse partition (small k) which is then refined in the subsequent tree levels. In addition the tree structure allows an efficient traversal through the tree to find the nearest feature. Therefore only $\mathcal{O}(\log(|\mathcal{V}|)$ vector comparisons are necessary instead of $\mathcal{O}(|\mathcal{V}|)$ comparisons when using a plain list (See Figure 1). The disadvantage of using the hierarchical approximated partitions instead of more accurate partitions produced by a plain k-means clustering is often made up for by the ability to use a larger vocabulary.

We would like to note that even though vocabulary trees are a more efficient way to cluster features than with non-hierarchical approaches, creating a vocabulary tree of more than 90,000 visual words is still a highly time-consuming task. However, the results of our experiments give rise to expectations that such a costly vocabulary creation needs to be done only once, and thus even more expensive but also more accurate clustering methods may be used in the future as the vocabularies can be re-used by many applications.

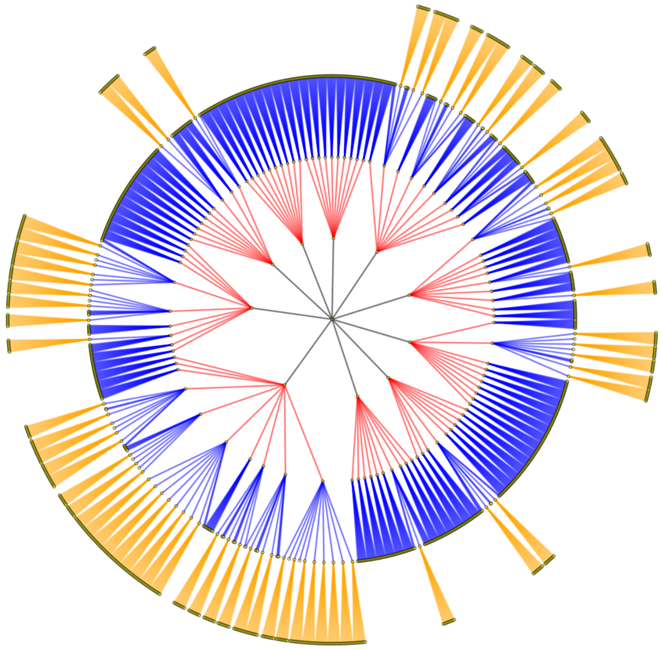

Fig. 2. An extreme example of a pruned vocabulary tree with $k = 10$ clusters per level and $l = 4$ levels. The actually constructed tree has 1594 leaves. The root node is at the center, every level of the tree is colored differently.

One difficulty in practice while building the vocabulary tree is to adjust the right number of sample vectors that are used to compute the tree. At each of l tree levels the current available clusters (the current leaves) are further subdivided by applying the k-means algorithm to each cluster. For each of these clusterings a sufficient number of sample vectors from within the respective cluster is required.

Once this number of vectors is below a minimum number the clustering is aborted, the respective tree branch stops and the cluster is regarded a leaf. Therefore, given a set of feature vectors, it is not known in advance how the resulting vocabulary tree's structure will look like. Unless the number of available feature vectors is very high it is likely that the tree will have some pruned branches and therefore less than k^l leaves. This is the reason why the number of distinct visual words of our visual vocabularies differs significantly from the targeted 100,000 words as the datasets we used for computing the different vocabularies to examine only have a limited number of images and therefore a limited number of feature vectors.

For illustration, Figure 2 shows one extreme example of such a pruned tree that was targeted to have 10,000 clusters by adjusting the number of cluster of each separate clustering with $k = 10$ and the number of levels to 4. Due to the lack of sample vectors during construction however, it only has 1594 leaves.

Of course one can always lower the minimum number of vectors used for clustering to sidestep this problem but this might degrade the accuracy of the computed feature space partitions.

5. PLSA IMAGE CLASSIFICATION

We apply pLSA image classification to various image domains in order to evaluate our visual vocabularies. We use pLSA models because they outperform basic bag-of-words models at image classification according to [5].

In this section we therefore briefly review the pLSA and how it is applied to images. Probabilistic latent semantic analysis (pLSA) was originally invented for text modeling where the document corpus consists of text documents. The goal of the pLSA is to assign mixtures of topics to texts in an unsupervised manner, i.e. the topics are not known in advance. The first step of this procedure is to build a term-document co-occurrence matrix of size $M \times N$, where M is the number of documents in the corpus and N denotes the number of words occurring in the documents all together. As the order of words is ignored by this matrix, this model is commonly known as a *bag-of-words* model.

We need to define equivalents to documents and words in order to be able to apply the pLSA to the image domain. Obviously, images can be considered documents. Words in the image domain are then called *visual words* and are derived as discussed in Section 4.3.

Having devised a visual vocabulary, we can replace extracted local features from all images of a given database by the most similar visual word, i.e. by the closest word in the feature vector space. Counting the occurring words in each image yields a term-frequency vector for each image. Together these vectors constitute the co-occurrence matrix for the image database. Given this co-occurrence matrix, the pLSA models the co-occurrence of visual words across images using a finite number K of hidden topics

The hidden topics usually model objects or object parts which regularly re-occur in the image database. Thus, an image can be explained as a mixture of multiple hidden topics and the occurrences of visual words within the image are assumed to be a result of this mixture. The probabilistic model is therefore:

$$P(d_i, w_j) = P(d_i) \sum_{K} P(z_k|d_i) P(w_j|z_k) \tag{1}$$

where $P(d_i)$ denotes the probability of document d_i to be drawn from the image database, $P(z_k|d_i)$ is the probability of the topic z_k given the current document, and $P(w_j|z_k)$ denotes the probability of the visual word w_j to occur given that it was caused by the topic z_k.

The latent topics cannot be observed directly, so we apply the Expectation-Maximization algorithm to model the topic distribution $P(z_k|d_i)$ for every image d_i from a database. We run the EM algorithm for 500 iterations in our experiments. The resulting topic distribution of an image is represented by a K-dimensional vector which constitutes a high-level representation of the image based on its visual features. At the same time it reduces the dimension of the image representation since we commonly choose the number of topics in our model to be significantly smaller than the number of visual words.

The topic vectors of a database of training images from given image classes can then be directly used for classifying query images [4]. We simply compute the topic vector of a query image and apply the k-nearest neighbor algorithm. That is, for the topic vector of the query image, we determine the k closest topic vectors among the topic vectors of the training images. The image class of the majority of the k nearest neighbors is considered the classification result.

6. EXPERIMENTAL RESULTS

We evaluate several visual vocabularies to examine if the origin (the underlying set of images) of visual vocabularies affects the classification performance of our bag-of-words-based image classification on different target domains. For each of the datasets discussed in Section 4.1 and for each feature type described in Section 4.2 three differently sized vocabularies are computed: A small vocabulary consisting of 500 visual words, a medium-sized vocabulary of about 9,000 to 10,000 visual words and a large-size vocabulary of about 90,000 to 100,000 visual words. Due to the reasons described in Section 4.3 the exact number of visual words varies depending on the actual dataset and its size.

Note that these experiments are neither meant to find the optimal vocabulary size nor the best parameter settings for the classification tasks or the pLSA model. Instead we compare the classification performances regarding the image sets the visual vocabularies are created from at several target domains.

Thus, we evaluate our visual vocabularies by examining the classification performance of pLSA models on four different image collections:

- 20,000 Flickr images and 2600 adult images
- 20,000 Flickr images and 2600 bikini images
- Caltech 101
- OT dataset

As the initial question has been raised during the development of an adult-image classification system, we test the ability of our classification system to discriminate adult images from non-adult images using both domain-specific vocabularies and vocabularies derived from other image sets. We further evaluate the classification performance of discriminating bikini images from non-bikini images.

We also evaluate the performance of discriminating all of the classes of the Caltech 101 dataset. This task is more difficult than separating the former image sets into two classes.

Finally we determine the classification performance of the different vocabularies on the OT dataset. The OT dataset [11] consists of 2688 images from eight scene categories: Coast,

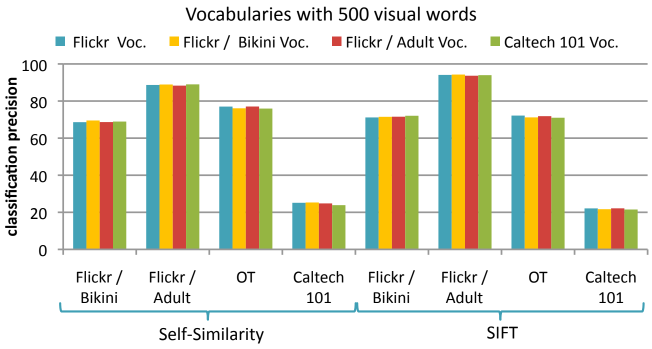

Fig. 3. Classification results based on visual vocabularies consisting of **500** words

Forest, Highway, Inside of Cities, Mountain, Open Country, Streets, Tall Buildings. This dataset was added to have a second score for discriminating images into more than two classes such as 'adult/non-adult' and 'bikini/non-bikini'.

Given the vocabularies and the classification task on a specified dataset we extract local features from each image and quantize them into discrete visual words. We then compute pLSA models for the derived co-occurrence tables. We use one half of the images of each category of the respective image set for training, i.e. for computing the pLSA model and the topic distributions as described in section 5. The remaining half is then used for testing.

Thus we compute the topic vectors for these images and classify them according to their k-nearest neighbors within the training set in topic vector space. An image of the test set is then classified by determining the majority of classes of the 9-nearest neighbors within the training set. That way we obtain the percentage of correctly classified images for each class. The overall classification performance on a given dataset is then computed as the average of these ratios. The resulting scores are much lower than computing the percentage of correctly classified images directly for the whole image set, but prevent the result to be dominated by classes with a large number of images.

Figures 3 to 6 show the evaluation results. In each figure, the vertical axis represents the classification score. On the horizontal axis, the four image sets are listed for both Self-Similarity and SIFT features. Each of the four grouped bars represents a different visual vocabulary, that is a visual vocabulary computed from a different set of images.

As can be seen, none of the vocabularies consisting of 500 words yielded a pLSA model which classified significantly better than the others. All models have almost identical average classification performances. We exemplarily show the classification performance split according to the image classes of the respective image sets in Figure 4. There are some variations for a few image classes such as the Highway class of the OT dataset, however these variations are compensated by the results for other classes.

Surprisingly, computing the vocabulary from image classes which are used for both training the pLSA model and testing it on unknown images does not improve the classification result. For example, one would expect that a model based on a vocabulary computed from adult and Flickr images should perform better in discriminating these two classes than models derived from other vocabularies. However, each of the other pLSA models classify these image classes equally well. Altogether the sets of images used for the computation of the vocabularies have no significant influence on classification performance.

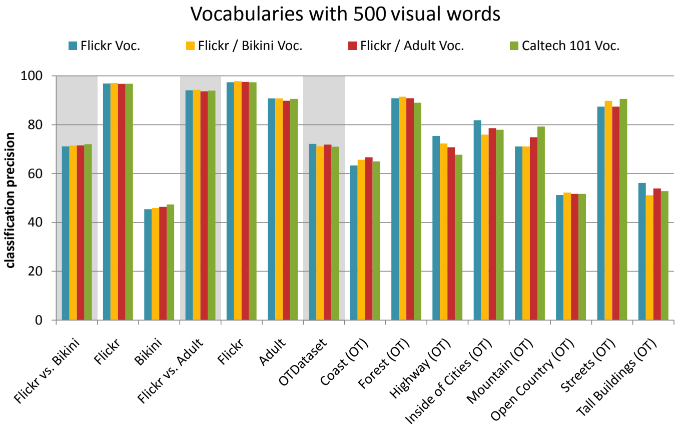

Fig. 4. Example results for all classes of the test datasets (except Caltech 101). The respective averaged results are shown on gray background. Vocabularies of 500 words were used for this figure. The results are analogous for larger vocabularies.

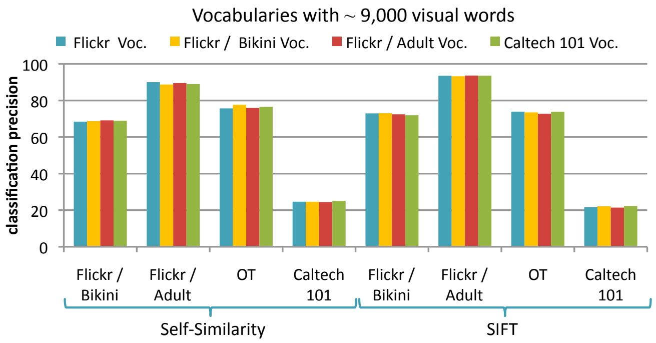

Fig. 5. Classification results based on visual vocabularies consisting of about **9,000** words

The same observations can be made for vocabularies consisting of more than 9,000 visual words as shown in Figure 5. All vocabularies yield similar classification performances. Furthermore, comparing these results to the ones from Figure 3 reveals that the performance of vocabularies consisting of 500 and 9000+ words respectively yield pLSA models with almost identical classification performances.

We also created visual vocabularies consisting of more than 90,000 words. Still, the origins of the visual vocabularies do not influence the classification results. The results of pLSA models based on these large vocabularies are also comparable to the results of the pLSA models with significantly smaller underlying vocabularies of 9000+ and 500 words. In fact, the differences

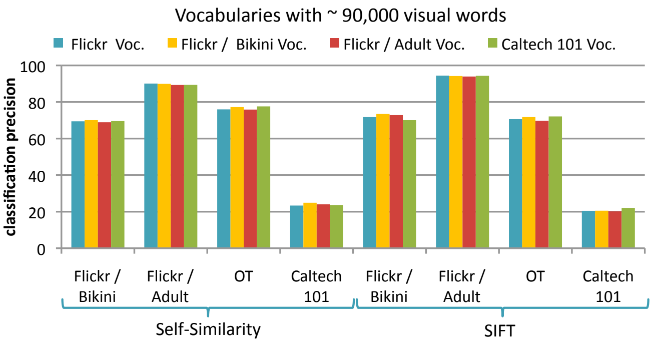

Fig. 6. Classification results based on visual vocabularies consisting of about **90,000** words

between the results depicted in Figures 6 and 5 are smaller than 1% for all four classification tasks.

In addition, as can be seen from Figures 3 to 6 all four vocabularies of each feature type perform equally well. Therefore, the above observations are also independent of the feature type, despite the fact that the performance on certain datasets varies when Self-Similarity or SIFT features are interchanged.

SIFT features apparently work better at distinguishing Flickr images from images with adult content and from bikini images by up to 5% and 3% respectively. On the other hand, the results on Caltech 101 and the OT dataset are improved by about 3% and 5% respectively when Self-Similarity features are used instead of SIFT.

7. CONCLUSION

Our experiments show that the origin of a visual vocabulary is neglectable when creating a bag-of-words model for image classification. The classification performance of such models does apparently not depend on the image set the underlying vocabulary was computed from. This observation also holds for various vocabulary sizes and feature types.

In other words, bag-of-words models do not require a specific, distinct visual vocabulary. This property presumably arises from the fact that images normally do not consist of random pixel arrangements but of a limited number of re-occurring structures and shapes. That is, the visual words computed from an arbitrary but diverse set of images are neither entirely random nor unique. As long as the number of images is sufficiently large, the resulting words are not biased towards special images sets.

This key insight allows to derive general universal visual vocabularies once and then reuse them for various applications without further refinement. Thus, the time needed for creating bag-of-words models can be reduced significantly since the time-consuming step of building the vocabulary needs to be done only once.

8. REFERENCES

[1] D. Nister and H. Stewenius, “Scalable recognition with a vocabulary tree,” *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2006 (CVPR 2006)*, vol. 2, pp. 2161–2168, 2006.

[2] T. Hofmann, “Probabilistic latent semantic analysis,” *Proc. of Uncertainty in Artificial Intelligence*, pp. 289–296, 1999.

[3] T. Hofmann, “Probabilistic latent semantic indexing,” *SIGIR 99: Proceedings of the 22nd annual international ACM Special Interest Group on Information Retrieval Conference on Research and development in information retrieval (SIGIR 1999)*, pp. 50–57, 1999.

[4] Anna Bosch, Andrew Zisserman, and Xavier Muñoz, “Scene classification via pLSA,” in *In Proc. European Conference on Computer Vision 2006 (ECCV 2006)*, 2006, pp. 517–530.

[5] R. Lienhart and M. Slaney, “pLSA on large scale image databases,” *IEEE International Conference on Acoustics, Speech and Signal Processing 2007 (ICASSP 2007)*, vol. IV, pp. 1217–1220, 2007.

[6] Eva Hörster, Thomas Greif, Rainer Lienhart, and Malcolm Slaney, “Comparing local feature descriptors in pLSA-based image models,” *30th Annual Symposium of the German Association for Pattern Recognition (DAGM 2008), Munich, Germany*, 2008.

[7] Eli Shechtman and Michal Irani, “Matching local self-similarities across images and videos,” *IEEE Conference on Computer Vision and Pattern Recognition 2007 (CVPR 2007)*, 2007.

[8] David G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] R. Fergus L. Fei-Fei and P. Perona, “Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories,” *IEEE. Conference on Computer Vision and Pattern Recognition 2004 (CVPR 2004), Workshop on Generative-Model Based Vision*, 2004.

[10] J. B. MacQueen, “Some methods for classification and analysis of multivariate observations,” *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, pp. 281–297, 1967.

[11] A. Oliva and A. Torralba, “Modeling the shape of the scene: A holistic representation of the spatial envelope,” *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.