Randomized Probabilistic Latent Semantic Analysis for Scene Recognition

Erik Rodner and Joachim Denzler

Chair for Computer Vision
Friedrich Schiller University of Jena
{Erik.Rodner,Joachim.Denzler}@uni-jena.de
http://www.inf-cv.uni-jena.de

Abstract. The concept of probabilistic Latent Semantic Analysis (pLSA) has gained much interest as a tool for feature transformation in image categorization and scene recognition scenarios. However, a major issue of this technique is overfitting. Therefore, we propose to use an ensemble of pLSA models which are trained using random fractions of the training data. We analyze empirically the influence of the degree of randomization and the size of the ensemble on the overall classification performance of a scene recognition task. A thoughtful evaluation shows the benefits of this approach compared to a single pLSA model.

1 Introduction

Building robust feature representations is an important step of many approaches to object recognition. Feature transformation techniques, such as principal component analysis (PCA) or linear discriminant analysis (LDA) offer the possibility to reduce the dimension of an initial feature space using a transformation estimated from all training examples. The main benefit is a compact representation, which exploits that feature vectors in high-dimensional spaces often lie on a lower dimensional manifold.

Within the typical bag-of-features (BoF) approach to image categorization, the reduction of feature vectors using probabilistic Latent Semantic Analysis (pLSA) showed to be beneficial for the overall classification performance [1,2]. The pLSA approach [3] originates from a text categorization scenario, in which a document is represented as an orderless collection of words. With pLSA the representation can be reduced to a collection of latent topics which generate all words of a document. It is natural to transfer this idea to an image categorization scenario and describe an image as a collection or bag of visual words [2]. An estimated distribution of visual word occurrences can be compressed into an image specific distribution of topics. As argued by [4], the pLSA approach has severe overfitting issues. This is due to the number of parameters, which increases with the number of training examples.

In this work, we describe a technique which prevents overfitting by building an ensemble of randomized subspaces and which significantly increase the

E. Bayro-Corrochano and J.-O. Eklundh (Eds.): CIARP 2009, LNCS 5856, pp. 945–953, 2009.

robustness and discriminative power of pLSA reduced features. The basic concept is similar to the random subspace methods of Ho [5] and Rodriguez et al. [6]. Instead of generating an ensemble of classifiers, our approach builds an ensemble of pLSA models which are used for feature transformation. This idea is related to multiple pLSA models used in Brants et al. [7]. Their approach exploits the diversity of generated models due to different random initializations of the EM algorithm which is used to estimate a model. In contrast to that, we generate multiple diverse feature transformations by utilizing the basic idea of Bagging [8] and train each model using a random fraction of the whole data.

Our method can directly be used for the application of scene recognition as described in Bosch et al. [2]. The goal is to categorize an image into a set of predefined scene types, such as mountain, coast, street and forest. Due to the high intraclass variation and low interclass distance, visual words tend to form groups of equal semantic meaning, which can be estimated using pLSA.

The remainder of this paper is structured as follows: The pLSA model and its connections to other approaches are described in Sect. 2. Section 3 presents and discusses our method of generating pLSA-models using a randomization technique. Experimental results within a scene recognition scenario are evaluated in Sect. 4 and show the benefits of our approach. A summary of our findings conclude the paper.

2 Probabilistic Latent Semantic Analysis

A standard approach to image categorization is the bag-of-features (BoF) idea. It is based on the orderless collection of local features extracted from an image and a quantization of these features into V visual words w_j, which build up a visual vocabulary. Images $\{d_i\}_i$ can be represented as a set of histograms $\{c_{ji}\}_i$ which counts how often a visual word w_j is the best description of a local feature in a specific image d_i [2]. Therefore this raw global feature vector associated with an image has as many entries as elements in the visual vocabulary. Especially in the context of scene recognition, it has been shown that the dimensionality reduction of BoF histograms using probabilistic Latent Semantic Analysis (pLSA) leads to performance benefits.

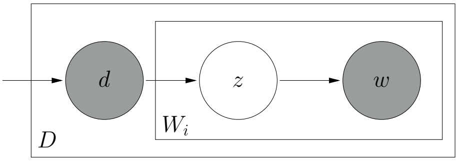


Fig. 1. The asymmetric model of probabilistic Latent Semantic Analysis (pLSA) in plate notation: (observable) visual words w are generated from latent topics z which are specific for each image d (W_i number of visual words, D number of images).

2.1 pLSA Model

The pLSA model, as shown in Fig. 1, models word and image (document) co-occurrences c_{ji} using the joint probability $p(w_j, d_i)$ of a word w_j and an image d_i in the following way:

$$p(w_j, d_i) = p(d_i) \sum_{k=1}^{Z} p(w_j \mid z_k)\, p(z_k \mid d_i). \tag{1}$$

For the sake of brevity, we use the same notation principles as in the original work [3], which abbreviates the event $\mathcal{W} = w_j$ with w_j and skips the formal definition of the random variables $\mathcal{W}, \mathcal{Z}$ and $\mathcal{D}$. Equation (1) illustrates that the pLSA model introduces a latent topic variable $\mathcal{Z}$ and describes all training images as a collection of underlying topics z_k. Note that this model is unsupervised and does not use image labels. By modeling all involved distributions as multinomial, it is possible to directly apply the EM principle to estimate them using visual word counts c_{ji} [3]. Additionally, we can rewrite (1) in matrix notation using $\mathbf{H} = [p(w_j, d_i)]_{j,i}$, $\mathbf{T} = [p(z_k \mid d_i)]_{k,i}$, $\mathbf{M} = [p(w_j \mid z_k)]_{j,k}$ and the diagonal matrix $\mathbf{D} = [p(d_i)]_{ii}$, which yields:

$$\mathbf{H} = \mathbf{M} \cdot \mathbf{T} \cdot \mathbf{D} \ . \tag{2}$$

This suggests a strong relationship to non-negative matrix factorization (NMF) as introduced by Lee and Seung [9]. In fact, it was highlighted by [10], that NMF of observed values $H_{ji} = c_{ji} \left(\sum_{j'i'} c_{j'i'} \right)^{-1}$ with Kullback-Leibler divergence is equivalent to the pLSA formulation which leads to an instance of the EM principle. In the subsequent sections, we will refer to the matrix $\mathbf{M}$ of topic-specific word probabilities as pLSA model, because it represents the image independent knowledge estimated from the training data.

2.2 pLSA as a Feature Transformation Technique

In [2], the pLSA technique is used as a feature transformation technique, similar to the typical application of PCA. The whole model can be seen as a transformation of BoF histograms $\mathbf{h}^i = [H_{ji}]_j$ into a new compact Z-dimensional description of each image as a vector of topic probabilities $\mathbf{t}^i = [p(z_k|d_i)]_k$.

Given an image with an unnormalized BoF histogram $\mathbf{h}$ that is not part of the training set, a suitable feature vector $\mathbf{t}$ has to be found. With a single image, the model equation (2) reduces to $\mathbf{h} = \mathbf{M}\mathbf{t}$ and the estimation of $\mathbf{t}$ can be done by applying the same EM algorithm used for model estimation but without reestimation of the pLSA model (matrix) $\mathbf{M}$. This idea is known as fold-in technique [3] and equivalent to the following NMF-optimization problem:

$$\mathbf{t}(\mathbf{M}, \mathbf{h}) = \operatorname*{argmin}_{\mathbf{t}'} \ \mathrm{KL}(\tilde{\mathbf{h}}, \mathbf{M}\mathbf{t}') \text{ w.r.t. to } \sum_k t'_k = 1 \ \ , \tag{3}$$

using the normalized BoF histogram $\tilde{\mathbf{h}} = \left(\sum_j h_j \right)^{-1} \mathbf{h}$ and the Kullback-Leibler divergence $\mathrm{KL}(\cdot, \cdot)$.

3 Randomized pLSA

As pointed out by Blei et al. [4], the estimation of the pLSA model leads to overfitting problems. This can be seen by considering the number of parameters involved which grows linearly with the number of training examples. A solution would be to use Latent Dirichlet Allocation [4] which demands sophisticated optimization techniques. In contrast to that, we propose to use an ensemble build by a randomization technique to solve this issue. As opposed to [7], which exploits the diversity of pLSA models resulting from random initializations of the EM-algorithm, we use a randomized selection of training examples, similar to the idea of Random Forests [8] and Random Subspaces [5].

Let $\{\mathbf{M}^l\}_{l=1}^M$ be an ensemble of pLSA models $\mathbf{M}^l = \mathbf{M}(\mathcal{T}^l)$ estimated using a random fraction $\mathcal{T}^l$ of the training data $\mathcal{T}$. We do not select training examples $(\mathbf{h}, y) \in \mathbb{R}^V \times \{1, \ldots, \Omega\}$ of a classification task with Ω classes individually. Instead we propose to select a random fraction of classes $\mathcal{C}^l \subset \{1, \ldots, \Omega\}$ with $|\mathcal{C}^l| = N$ and use all training examples $\mathcal{T}^l = \bigcup_{y_i \in \mathcal{C}^l} \{\mathbf{h}^i\}$ of each selected class . This allows estimating topics which are shared only among a subset of all classes. Each pLSA model $\mathbf{M}^l$ is used to transform BoF histograms $\mathbf{h}^i$ into topic distributions $\mathbf{t}(\mathbf{M}^l, \mathbf{h}^i)$. For training examples in $\mathcal{T}^l$, we use the topic distributions resulting from the pLSA model estimation. All other training examples and each test example are transformed using the "fold-in" technique defined by (3).

One commonly used technique to combine feature transformation models is simply averaging outputs [5] of classifiers trained for each feature set individually. This technique does not allow the classifier to learn dependencies between different models. Therefore we use a concatenation of all calculated feature vectors $\mathbf{t}(\mathbf{M}^l, \mathbf{h}^i)$ as a final feature $\mathbf{t}(\mathbf{h}^i)$:

$$\mathbf{t}(\mathbf{h}^i)^T = \left(\mathbf{t}(\mathbf{M}^1, \mathbf{h}^i)^T, \ldots, \mathbf{t}(\mathbf{M}^M, \mathbf{h}^i)^T\right) \quad . \tag{4}$$

These final feature vectors are of size $M \cdot Z$ and can be used to train an arbitrary classifier. In our experiments, we use an one-vs.-one SVM classifier with a radial basis function kernel.

We have to estimate M pLSA models with the EM algorithm, thus we need roughly M times the computation time of a single model fit. To be exact, we use a fraction of the training data for each model estimation and have to perform the EM algorithm with the "fold-in" technique for each remaining training example:

$$\text{time}_{\text{randomized-plsa}} = \sum_{l=1}^{M} \left(\frac{|\mathcal{T}^l|}{|\mathcal{T}|} \text{time}_{\text{single-model}} + \left(|\mathcal{T}| - |\mathcal{T}^l|\right) \text{time}_{\text{fold-in}} \right) \quad . \tag{5}$$

Therefore we pay for the advantage of reduced overfitting with a higher computational cost.

4 Experiments

We experimentally evaluated our approach to illustrate the benefits of randomized ensembles of pLSA models. In the following, we empirically validate the following hypotheses:

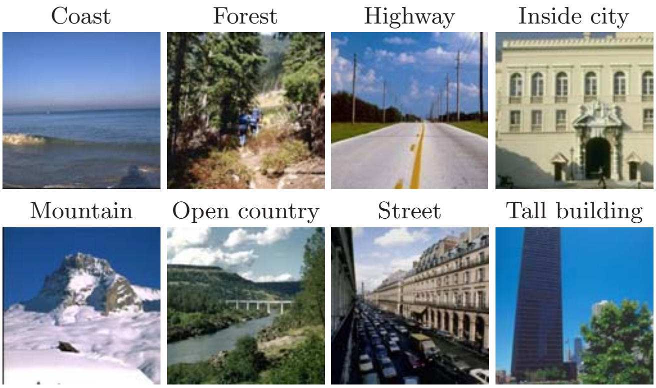


Fig. 2. Example images of each class of the dataset of [11] which we use for evaluation

1. Randomized pLSA ensembles lead to a performance gain in comparision to single pLSA and the usual BoF method, which is most prevalent with a large set of training examples. (Sect. 4.2)
2. With an increasing size M of the ensemble, the recognition rate increases and levels out after a specific size. (Sect. 4.3)
3. The optimal selection of the parameter N (size of the random subset of classes) depends on the size of the training set. (Sect. 4.2)

Additionally, in contrast to other researchers [2], we found that the single pLSA method, in general, does not result in significantly better performance compared to the standard BoF method. A discussion and detailed results of our experiments can be found in Sect. 4.2.

4.1 Experimental Setup

The analysis of the benefits and involved parameters of our method is done using the performance evaluation within a scene recognition scenario. To evaluate our randomized pLSA technique, we use the image dataset of Oliva and Torralba [11], which is a publicly available set of images for evaluating scene recognition approaches [2]. It consists of images from eight different classes which are shown exemplarily in Fig. 2.

All color images are preprocessed as described in [2]. The performance of the overall classification system is measured using unbiased average recognition rates. In contrast to previous work [2], we use Monte Carlo analysis by performing ten independent training and test runs with a randomly chosen training set. This provides us with a statistical meaningful estimate and allows to compare three different approaches: (1) standard BoF without pLSA using normalized histograms (BoF-SVM), (2) a single pLSA model (pLSA) and (3) an ensemble with a varying number of pLSA models (r-pLSA). For the BoF approach directly using BoF histograms **h** as feature vectors, we applied thresholding using mutual information (MI) [12] resulting in a performance gain of 5% for this case.

In all experiments, the number of topics Z is set to 25 and a vocabulary of 1500 visual words is created using the method described in Sect. 4.1. The influence of these parameters was analyzed in previous work [2] and the values showed to be optimal for the dataset of [11].

Feature Extraction. As a local feature representation, we use the OpponentSIFT method proposed in [13]. The task of scene recognition requires the use of information from all parts of the image. Therefore, local descriptors are calculated on a regular grid rather than on interest points only.

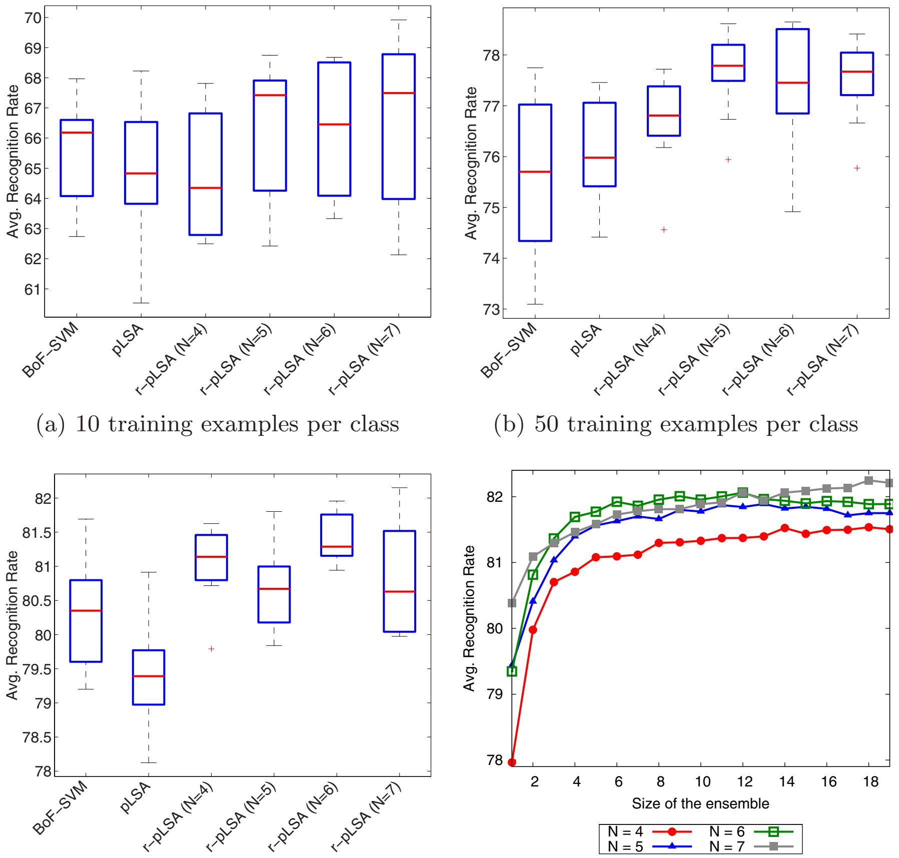
The method of [12], which utilizes a random forest as a clustering mechanism, is used to construct the codebook. It trains a random forest classifier using all local features and corresponding image labels. The leafs of the forest can then be interpreted as individual clusters or visual words. This codebook generation procedure showed superior results compared to standard k-means within all experiments. It also allows us to create large codebooks in a few minutes on a standard personal computer. Note that due to the ensemble of trees, this approach results in multiple visual words for a single local feature. This is not directly modeled by the graphical model underlying pLSA as can be seen in Fig. 1. Nevertheless we can still apply pLSA on the resulting BoF histograms.

4.2 Results and Evaluation

For a different number of training examples (for each class), Figures 3(a) - 3(c) show a comparision of our approach using randomized pLSA ensembles with a standard BoF approach and the utilization of a single pLSA model [2], which is equivalent to randomized pLSA with $N = 8$ and $M = 1$. The classification rates of our approach are displayed for different values of N. To display the results of the multiple training and test runs, we use box plots [14].

At first it can be seen that for nearly all settings (except for 10 training examples and $N = 4$), our randomized pLSA method reaches a higher recognition rate than the usual BoF approach and the method using a single pLSA model [2]. These performance benefits are most prevalent with a large number of training examples. Another surprising fact is that the method proposed by [2] is not significantly better than the simple BoF method. This might be due to our use of MI-thresholding for raw BoF histograms. Another reason could be the analysis using fixed training and test sets in the comparision performed by [2], which does not lead to significant results. With a glance at the box plots for different values of N, we can see that it is hard to determine an optimal parameter value. However a value of $N = 5$ seems to be a reasonable choice.

Note that the absolute recognition performance of $81 - 82\%$ for 150 examples is lower than the best values obtained by [2], which are 87.8% on a test set and 91.1% on a validation set. This is mainly due to different local features and the incorporation of spatial information, which we do not investigate in this paper. However, our idea of randomized pLSA ensembles could be well adopted to use spatial pyramids as proposed in [2].



(a) 10 training examples per class

(b) 50 training examples per class

(c) 150 training examples per class

(d) Influence of the ensemble size

Fig. 3. Evaluation using average recognition rate of the whole classification task: (a-c) Comparision of a usual BoF approach (BoF-SVM), pLSA reduced features and our approach utilizing a randomized ensemble of multiple pLSA models (r-pLSA) using training examples from $N = 4, 5, 6, 7$ random classes. The median of the values is shown by the central mark, top and bottom of the box are the 0.25 and 0.75 percentiles, the whiskers extend to the maximum and minimum values disregarding outliers, and outliers are plotted individually by small crosses [14]. 3(d) classification performance of r-pLSA with a varying size of the ensemble for a fixed training and test set.

4.3 Influence of the Ensemble Size

As can be seen from Fig. 3(d), increasing the number M of pLSA models yields a better overall performance. As expected this leads to convergence after a specific size of the ensemble. A similar effect of the ensemble size can be observed when using Random Forests [8]. Because of the ability of the SVM classifier to build maximum margin hypotheses, the effect of overfitting due to an increasing number of features, and thus to an increasing VC dimension, does not occur.

5 Conclusion and Further Work

We showed that utilizing a randomization principle, an ensemble of pLSA models can be build, which offers a feature transformation technique that is not prone to overfitting compared to a single pLSA model. In a scene recognition scenario, this technique leads to a better recognition performance in comparision with a single model or a standard bag-of-features approach. Our experiments also showed that the recognition performance increases with more pLSA models and levels out. An interesting possibility for future research would be to study ensembles of models estimated with Latent Dirichlet Allocation, which is a more sophisticated method for topic discovery and a well-known Bayesian method [4]. Finally, experiments should be performed using other datasets with more classes and analyzing the trade-off between a better recognition rate and a higher computational cost.

References

1. Quelhas, P., Monay, F., Odobez, J.M., Gatica-Perez, D., Tuytelaars, T., Van Gool, L.: Modeling scenes with local descriptors and latent aspects. In: Proceedings of the Tenth IEEE International Conference on Computer Vision, pp. 883–890 (2005)
2. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. IEEE Trans. Pattern Anal. Mach. Intell. 30(4), 712–727 (2008)
3. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1-2), 177–196 (2001)
4. Blei, D., Ng, A., Jordan, M.: Latent dirichlet allocation. The Journal of Machine Learning Research 3, 993–1022 (2003)
5. Ho, T.K.: The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20(8), 832–844 (1998)
6. Rodriguez, J.J., Kuncheva, L.I., Alonso, C.J.: Rotation forest: A new classifier ensemble method. IEEE Trans. Pattern Anal. Mach. Intell. 28(10), 1619–1630 (2006)
7. Brants, T., Chen, F., Tsochantaridis, I.: Topic-based document segmentation with probabilistic latent semantic analysis. In: Proceedings of the Eleventh International Conference on Information and Knowledge Management, pp. 211–218 (2002)
8. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001)
9. Lee, D., Seung, H.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, vol. 1998, pp. 556–562. MIT Press, Cambridge (2001)
10. Gaussier, E., Goutte, C.: Relation between plsa and nmf and implications. In: Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 601–602 (2005)

11. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vision 42(3), 145–175 (2001)
12. Moosmann, F., Triggs, B., Jurie, F.: Fast discriminative visual codebooks using randomized clustering forests. In: Advances in Neural Information Processing Systems, pp. 985–992 (2006)
13. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluation of color descriptors for object and scene recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
14. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley, Reading (1977)