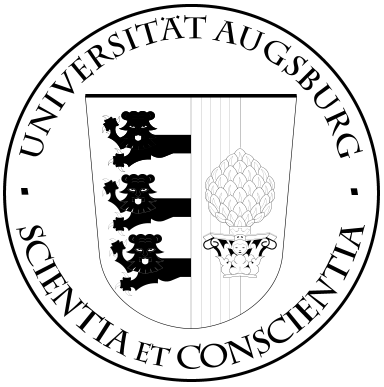
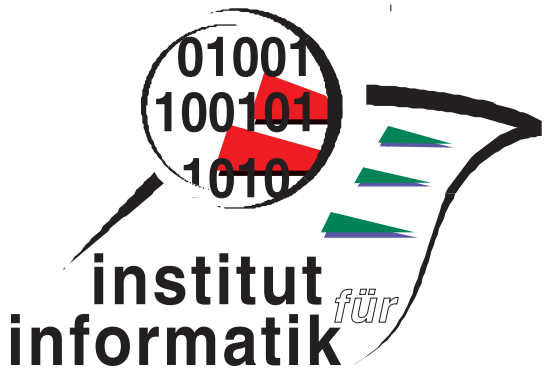
Universität Augsburg



Continuous Visual Vocabulary Models for pLSA-Based Scene Recognition

E. Hörster, R. Lienhart, M. Slaney

Report 2008-05 April 2008



Institut für Informatik
D-86135 Augsburg

Continuous Visual Vocabulary Models for pLSA-Based Scene Recognition

Eva Hörster
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
hoerster@informatik.uni-augsburg.de

Rainer Lienhart
Multimedia Computing Lab
University of Augsburg
Augsburg, Germany
lienhart@informatik.uni-augsburg.de

Malcolm Slaney
Yahoo! Research
Santa Clara, CA
USA
malcolm@ieee.org

ABSTRACT

Topic models such as probabilistic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation (LDA) have been shown to perform well in various image content analysis tasks. However, due to the origin of these models from the text domain, almost all prior work uses discrete vocabularies even when applied in the image domain. Thus in these works the continuous local features used to describe an image need to be quantized to fit the model. In this work we will propose and evaluate three different extensions to the pLSA framework so that words are modeled as continuous feature vector distributions rather than crudely quantized high-dimensional descriptors. The performance of these continuous vocabulary models are compared in an automatic scene recognition task. Our experiments clearly show that the continuous approaches outperform the standard pLSA model. In this paper all required equations for parameter estimation and inference are given for each of the three models.

1. INTRODUCTION

Scene recognition or scene classification is the task of automatically assigning an image to one category out of a fixed number of scene categories. A closely related task is image retrieval, which consists of finding images of similar content to a candidate image. Both tasks become more important as personal as well as on-line image repositories grow. In this work we will focus on a scene recognition task. Nevertheless the proposed approaches may also be applied to other tasks such as image retrieval.

Probabilistic models with hidden/latent topic variables such as probabilistic Latent Semantic Analysis (pLSA) [8] and Latent Dirichlet Allocation (LDA) [4] and their extensions are popular in the document and language modeling community as well as in the pattern recognition community. Originally developed for the purpose of text document modeling in large collections, those models have been introduced and re-purposed for image content analysis tasks including object recognition [16], scene recognition [11], automatic image segmentation and image annotation [2].

Latent topic models model each document in a collection as a distribution over a fixed number of topics. Each topic aims to model the co-occurrence of words inside and across the documents and is in turn characterized by a distribution over a fixed size and discrete vocabulary. Applied to visual tasks, the distribution of hidden topics in an image refers to the degree to which an abstract object such as grass, water, sky, street, etc. is contained in the image. This gives rise to a low-dimensional description of the coarse image content, which can be used e.g. to enable classification or retrieval of images.

As these models have originally been designed for text analysis, words are modeled as discrete variables. In the visual domain we are challenged with the fact that visual features describing an image, especially local image descriptors, are often continuously distributed in some high dimensional space. Thus visual features are quantized into a fixed-size visual vocabulary in order to be able to apply the original pLSA model to image analysis tasks.

In most related efforts, quantization is done by clustering descriptor vectors, representing each cluster by one visual feature vector (the so-called "cluster center"), and subsequently mapping each feature vector to its closest cluster center in order to get a visual word representing the descriptor vector. However, this procedure does not necessarily produce optimal results since for example, it does not account for the distance of features to their closest cluster center.

In speech recognition applications it has been shown that introducing continuous variable models, especially in the case of Hidden-Markov-Models (HMMs), significantly improves performance [20]. Thus in this work we introduce and study models in which continuous visual vocabulary models are considered and thus we model words with continuous, high-dimensional feature vector distributions. We propose three different approaches that extend the discrete pLSA model. We will competitively evaluate their performance in a scene recognition task using the results from a discrete pLSA model as the baseline. Parameter estimation and inference algorithms are presented for each of the proposed models.

1.1 Related Work

There exist basically two fundamental approaches to solve the scene recognition task so far. Earlier works [17, 18] considered mostly global, low level image features to describe the scene images. It later work [19, 5, 11, 15, 14] intermediate concepts are used to represent the images. Whereas some models [14, 19] use supervision to learn the concepts, others [5, 11, 15] employ probabilistic latent topic models to learn the concepts in an unsupervised manner.

Latent topic models have been applied successfully in several image content analysis tasks such as object categorization [16], image retrieval [12, 9], automatic segmentation [6] and automatic annotation [2]. Further variations of latent space models have been applied to the problem of modeling annotated images [2, 3]. Especially in a scene classification task, as considered in this paper, these models have been shown by Bosch et al. [5] to outperform previous models [14, 19].

The majority of these related works use quantized local image descriptors as the starting point to build their model. As mentioned earlier the mapping from continuously distributed local features to a discrete visual vocabulary does not necessarily lead to optimal performance. In this work we therefore consider continuously vocabulary models which do not require a quantization of the high-dimensional feature vectors. In the context of latent topic models there has been very little work in this area. In order to model annotated data, Blei et al. used a multivariate Gaussian to represent image regions conditioned on a topic variable in two extensions of the LDA [3].

The two closest related works to our approach are the work by Ahrendt et al. [1] and the work of Larlus and Jurie [10]. The first work [1] proposes the so called Aspect Gaussian Mixture Model (AGMM), which extends the pLSA model to the case of continuous feature vectors. This model is equivalent to our second proposed model, the SGW-pLSA. The model is evaluated in a music genre classification task. However, they use supervised training with known concepts for each training sample, while we learn the model's parameters completely unsupervised. In the second work [10] a similar extension is proposed for the closely related LDA model. Gibbs sampling is used for parameter estimation, and the model is applied in an object categorization task. The difference to our work is that we propose three different models and we apply those models in a scene recognition task. Furthermore we consider the pLSA model instead of the LDA model and we perform parameter estimation via the EM algorithm.

1.2 Contributions

The main contributions of this paper are:

- We propose three different extensions to the pLSA which model the visual vocabulary continuously.
- We present the algorithms for parameter estimation and inference for each proposed model.
- We perform a competitive evaluation of the models in a scene recognition task taking the discrete pLSA model as the baseline.

The paper is organized as follows. Section 2 describes the baseline system we use to solve the scene recognition task. We discuss local feature generation as well as the visual vocabulary computation and we review the pLSA model in detail. We present our three proposed continuous vocabulary pLSA models in Section 3. In Section 4, we show how parameter estimation and inference is performed for each of the models. In Section 5 we describe the experimental evaluation, we show and discuss results. Finally we summarize and conclude the paper in Section 6.

2. PLSA-BASED SCENE RECOGNITION

In this work we adopt the scene recognition approach proposed by Bosch et al. [5]. This approach uses a discrete pLSA model to represent each image in a database. In the next section we will then present the continuous vocabulary pLSA models and explain how the scene recognition system is modified when using a continuous visual vocabulary pLSA model instead of the discrete one.

The pLSA [8] was originally developed in the context of text modeling, where words are the elementary parts of documents. Documents are modeled as mixtures of intermediate hidden topics, and topics are characterized by a distribution over words. Applied to image modeling, the images are our documents. The mixture of hidden topics then refers to the degree to which certain objects or certain object parts are contained in an image. It is important to note that pLSA allows us explicitly to represent an image as a mixture of topics, i.e. as a mixture of one or more objects/object parts. Since for all currently practical applications the number of topics modeled is much smaller than the number of visual words, the topic distribution gives rise to a compact, low-dimensional description of the image content.

The starting point for building a pLSA model is to represent the entire corpus of documents by a term-document co-occurrence matrix of size $M \times N$. M indicates the number of documents in the corpus and N the number of different words occurring across the corpus. Each matrix entry contains the number of times a specific word (column index) is observed in a given document (row index). Such a representation ignores the order of words in a document, and is commonly called a *bag-of-words model*.

In order to construct a co-occurrence table in the visual domain, and thus to be able to apply the pLSA model unchanged, we first need to define an equivalent to words in documents. These elementary parts are commonly called visual words. They are usually derived by vector quantizing automatically extracted local image descriptors.

In our work, visual words are derived by clustering a subset of automatically extracting local image descriptors from the training images using the k-means algorithm. The means of each cluster are kept as the visual words, together they form the visual vocabulary.

Given the vocabulary, features are extracted from each image in the database. Each image d_i is represented as consisting of N_i visual words by replacing each detected feature vector by its most similar visual word, defined as the closest word in the high-dimensional feature space. The word oc-

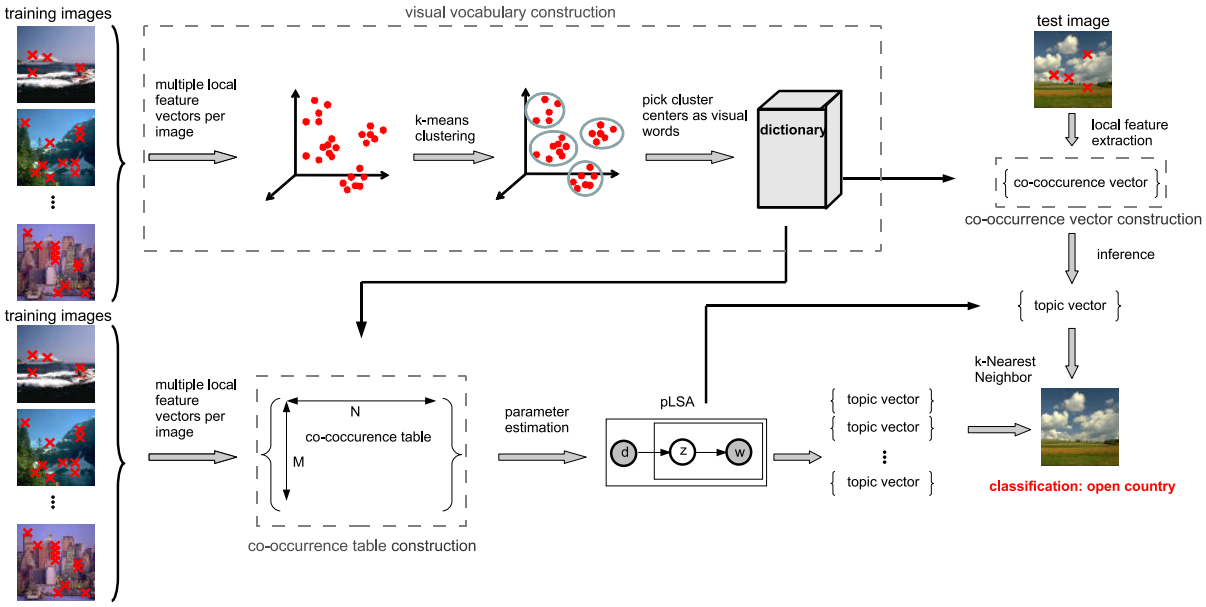


Figure 1: Scene classification system based on a discrete pLSA model.

currences are counted leading to the term-frequency vector for each image document and the term-frequency vectors of all images constitute then the co-occurrence matrix. Since the order of terms in a document is ignored, any geometric relationship between the occurrences of different visual words in images is disregarded.

Given the co-occurrence matrix, the pLSA uses a finite number of hidden topics to model the co-occurrence of visual words inside and across image document. This model assumes that every word occurring in a document in the corpus is associated with a hidden, i.e. unobservable, topic variable. Probability distributions of the visual words given a hidden topic as well as probability distributions of hidden topics given the documents are learned in an unsupervised fashion. We estimate the model parameters on the training data and apply this model to all images in the database in order to derive the topic distribution for each image. This low-dimensional topic vector is further used to represent each image document.

For scene recognition we perform a simple k-Nearest Neighbor (kNN) search for the topic vectors of the unlabeled test images over the labeled training images using the L2-norm as distance metric. More sophisticated distance metrics and/or machine learning algorithms such as Support Vector Machines (SVMs), Random Forrest (RF), or Adaboost could be applied for improving the recognition results further. As our main goal in this work is to compare the original discrete word model with three different continuous word models in the pLSA framework, we have chosen a simple kNN approach. Figure 1 gives a system overview.

2.1 Local Feature Descriptors

For this work we chose the well-known SIFT features proposed by David Lowe [13] as local image descriptors. They are computed in two steps: A sparse set of interest points

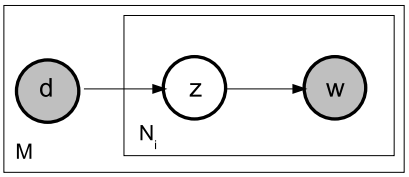


Figure 2: Graphical representation of the pLSA model. M denotes the number of images in the database and N_i the number of visual words in image d_i. Shaded nodes highlight the observable random variables w for the occurrence of a visual word and d for the respective document. z denotes the hidden topic variable.

is detected at extrema in the difference of Gaussian pyramid, and a scale and orientation are assigned to each interest point besides its position. Then we compute a 128-dimensional gradient-based feature vector from the local gray scale neighborhood of each interest point in a scale and orientation invariant manner.

Note that each image usually leads to a different number of features even if two images have the same size. The number of feature computed depends on the structure and texture of the image.

After having computed all 128-dimensional SIFT feature vectors for each image we perform a whitening PCA to extract the 75 most important components from the 128-dimensional vectors. This is done by only keeping the 75 components belonging to the largest eigenvalues. The lower dimensionality ensures faster computation of the pLSA models. Our experiments also showed that no/very little performance is lost due to this dimensionality reduction, compared to the original 128-dimensional feature vectors.

2.2 pLSA

We assume that in the discrete case each image d_i is represented as consisting of N_i visual words w_j. There are K different visual words in the vocabulary and each w_j is a 75 dimensional feature vector.

The pLSA model then assumes that the following generative process has created the co-occurrence matrix [8]:

- Pick a document d_i with prior probability $P(d_i)$
- Select a latent topic z_h with probability $P(z_h|d_i)$
- Generate a (visual) word w_j with probability $P(w_j|z_h)$

Note that the number of topics and words are predefined. This generative process results in the following model:

$$P(w_j, d_i) = P(d_i) \sum_{h=1}^{H} P(w_j|z_h)P(z_h|d_i) \quad (1)$$

where H denotes the number of topic in the model. Figure 2 shows the graphical representation of the pLSA model.

The probability distributions $P(w_j|z_h)$ of the visual words w_j given a hidden topic z_h as well as the probability distributions $P(z_h|d_i)$ of hidden topics z_h given the images d_i are learned completely unsupervised by means of the Expectation Maximization (EM) algorithm [7]. For detailed equations please refer to [8]. The topic distributions of new images that are not part of the original training corpus are estimated by a fold-in technique [8]. Here the EM algorithm is applied to the unseen images. However, this time the word distributions conditioned on the topic $P(w_j|z_h)$ are fixed (i.e., not updated) and only the topic distribution $P(z_h|d_i)$ for each image is computed.

3. APPROACHES

When applying the discrete pLSA model to image data, the high-dimensional feature vectors need to be quantized first in order to obtain a fixed number of discrete, visual words. However, the quantization procedure, as described in the previous section, is not necessarily optimal. In this work we will therefore describe three different ways to model directly the probability of features vectors under each topic, and thus making the quantization of the descriptors obsolete.

Ideally we would like to have a separate probability distribution over the feature space for each topic. We do this by using Gaussian Mixture Models (GMM). We call this model GM-pLSA and describe it in Section 3.3. But a model of this complexity is expensive to train, both in time and data. Thus we also test two simplifications that reduce the model complexity.

In a slightly simpler approach we learn Gaussians that are shared across all topics. In Section 3.1 we describe the SGW-pLSA model that learns the means and covariances of a single set of Gaussians as part of the topic determination algorithm.

A further computational simplification is possible if we cluster the feature data in advance, much as is done for discrete pLSA, and learn the probability of each cluster given a topic. We represent each cluster by a Gaussian distribution. This model is called FSGW-pLSA and is described in Section 3.2. Figure 3 shows an overview of the different model structures.

In the continuous case we represent each image d_i as consisting of N_i local feature descriptors f_j.

3.1 pLSA with Shared Gaussian Words (SGW-pLSA)

In the SGW-pLSA approach we modify the original pLSA model such that each word is represented by a multivariate Gaussian distribution and we assume that each high-dimensional feature vector is sampled from one of those Gaussian distributions. This results in modeling the topics, i.e. the probabilities $P(w|z)$, by a multivariate mixture of Gaussian distributions, where Gaussians are shared between the different topics.

This approach is similar to the model presented by Larlus et al. [10] for the case of the LDA – a pLSA related model. For the case of pLSA a similar model has been presented but the authors consider only supervised learning [1].

The SGW-pLSA model assumes the following process for sampling a feature descriptor f_j from the image database:

- Pick a document d_i with prior probability $P(d_i)$
- Select a latent topic z_h with probability $P(z_h|d_i)$
- Choose a Gaussian component g_k depending on the chosen topic z_h with probability $P(g_k|z_h)$
- Sample a descriptor f_j from $N(f_j|\mu_k, \Sigma_k)$, which is a multivariate Gaussian distribution over the feature vector space modeling the Gaussian component g_k

According to this generative process, equation 1 becomes:

$$P(f_j, d_i) = P(d_i) \sum_{h=1}^{H} \sum_{k=1}^{K} P(f_j|g_k) \cdot P(g_k|z_h) \cdot P(z_h|d_i) \quad (2)$$

where

$$P(f_j|g_k) = N(f_j|\mu_k, \Sigma_k). \quad (3)$$

Here H and K denote the total number of the topics and Gaussian words in the model, respectively.

It can be seen that the parameters of the Gaussian distributions, i.e. of the continuous visual vocabulary, become part of the model. Thus, those parameters are estimated simultaneously with the other model parameters in the learning algorithm (see Section 4.1). Additionally we can also omit the computation of the co-occurrence table/vector in our scene recognition system (see Figure 1).

As in the case of a discrete pLSA model the necessary quantization is performed before the actual model computation and thus does not account for the probabilistic model learned in the subsequent step, the joint learning of the Gaussian distributions with the other pLSA parameters may be advantageous. On the other hand the SGW-pLSA model estimation

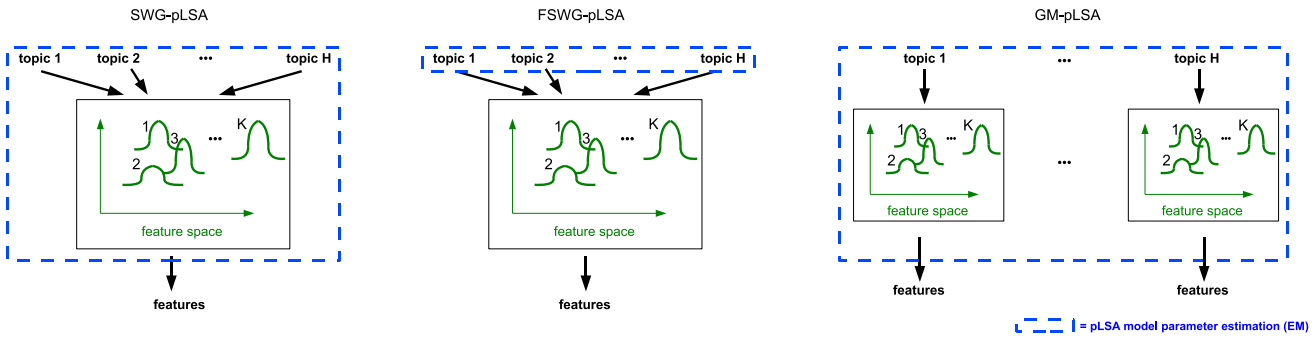


Figure 3: Model structure of the three proposed continuous vocabulary pLSA approaches.

might be more difficult as many more parameters (i.e., the means and covariance matrices) must be estimated.

3.2 pLSA with Fixed Shared Gaussian Words (FSGW-pLSA)

In order to examine the influence of modeling the visual words continuously by Gaussian distributions, we proposes the FSGW-pLSA. Here we assume the same probabilistic model as in the SGW-pLSA. However, during the model estimation we do not explicitly estimate the parameters of the Gaussian distributions representing the words.

We learn an ordinary Gaussian mixture model representing the shared continuous vocabulary on the extracted local image descriptors of the training image set in advance. Then, in the subsequent probabilistic model computation of the SGW-pLSA we assume the parameters of the Gaussians, i.e. the means μ_k and covariance matrices Σ_k, are fixed and only the topic and component probabilities $P(z_h|d_i)$ and $P(g_k|z_h)$ are estimated.

Summarizing, in the FSGW-pLSA, the words are modeled by a continuous distribution over the feature space, making quantization unnecessary. In contrast to the SGW-pLSA, the parameters are of the Gaussian distributions modeling the continuous visual vocabulary are computed separately previous to the topic model parameter estimation.

3.3 pLSA with Gaussian Mixtures (GM-pLSA)

In the above two approaches (SGW-pLSA and FSGW-pLSA) all topics share a single visual vocabulary. It may be beneficial to allow for different means and covariance matrices for each topic, i.e. no sharing of Gaussian components between topics. This results in modeling each topic, i.e. the probabilities $P(f_j|z_h)$, by its individual multivariate Gaussian mixture model over the feature space. Thus given a topic we select a Gaussian mixture component out of the Gaussian mixture model associated with the topic, and depending on the mixture component the feature is sampled.

We assume that each feature f_j in image d_i is generated as follows:

- Pick a document d_i with prior probability $P(d_i)$
- Select a latent topic z_h with probability $P(z_h|d_i)$
- Choose a Gaussian component g_k^h depending on the topic z_h with probability $P(g_k^h|z_h) = \pi_{hk}$, where g_k^h is the k-th Gaussian component associated with topic h
- Sample a descriptor f_j from $N(f_j|\mu_{kh}, \Sigma_{kh})$, which is a multivariate Gaussian distribution over the feature vector space modeling the multivariate Gaussian distribution g_k^h

According to this generative process, introducing a multivariate Gaussian mixture over the feature space for each topic z_h

$$P(f_j|z_h) = \sum_{k=1}^{K} \pi_{kh} \cdot N(f_j|\mu_{kh}, \Sigma_{kh}) \tag{4}$$

yields to the following model:

$$P(f_j, d_i) = P(d_i) \sum_{h=1}^{H} \left(P(z_h|d_i) \cdot \sum_{k=1}^{K} \pi_{kh} \cdot N(f_j|\mu_{kh}, \Sigma_{kh}) \right) \tag{5}$$

In contrast to the model described in the previous subsection, here the multivariate Gaussian distributions modeling the feature space are not shared, thus the means and covariances of the K Gaussians are different for each topic. On the one hand this enables to use the optimal means and covariances for each topic. On the other hand, as we need more Gaussians in total to model all topics, the number of parameters in the model is significantly larger for the same number of Gaussians per topic. Having observed that most topics are only represented by a small number of words/Gaussians compared to the entire number of visual words/Gaussians in the model, we should be able to reduce the number of Gaussians per topic without performance degradations. Thus, in our experiments we use fewer Gaussians to represent a topic compared to the SGW-pLSA and FSGW-pLSA approaches. However the total number of Gaussians and parameters for this third model will still be larger than the number for the other two models.

As in the SGW-pLSA model, in the GM-pLSA model the computation of the Gaussian distributions parameters and therefore the continuous visual vocabulary becomes part of

the model estimation. We can also omit the computation of the co-occurrence table/vector in our scene recognition model if we replace the discrete pLSA model by the GM-pLSA model (see Figure 1).

4. PARAMETER ESTIMATION

We will now present the algorithms for parameter estimation and inference in the three proposed continuous vocabulary pLSA models.

4.1 SGW-pLSA

According to the SGW-pLSA model (Equation 2 and 3), the log likelihood l of all images in the database is given by:

$$l = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \log \left(\sum_{h=1}^{H} \sum_{k=1}^{K} [P(d_i) \cdot P(z_h|d_i) \cdot P(g_k|z_h) \cdot N(f_j|\mu_k, \Sigma_k)] \right) \quad (6)$$

where M denotes the number of images in the database and N_i the number of local descriptors representing the image d_i.

During model estimation we need to learn the topic and component probabilities $P(z_h|d_i)$ and $P(g_k|z_h)$ as well as the parameters of the Gaussian distributions $N(\cdot|\mu_k, \Sigma_k)$. Due to the existence of the sums inside the logarithm, direct maximization of the log-likelihood by partial derivatives is difficult. Thus we use the Expectation Maximization (EM) algorithm [7]. The EM-algorithm is an iterative optimization method that alternates between two update steps. The expectation step (E-step) in the EM-algorithm consists of estimating the posterior probabilities for the latent variables taking as evidence the observed data and the current parameter estimates. Thus in the E-Step we calculate the variables β_{kh}[1].

$$\beta_{kh}^{ij} = \frac{P(z_h|d_i) \cdot P(g_k|z_h) \cdot N(f_j|\mu_k, \Sigma_k)}{\sum_{h=1}^{H} \sum_{k=1}^{K} P(z_h|d_i) \cdot P(g_k|z_h) \cdot N(f_j|\mu_k, \Sigma_k)} \quad (7)$$

The M-step consists of maximizing the expected complete data-likelihood $E(l^{comp})$:

$$E(l^{comp}) = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{h=1}^{H} \sum_{k=1}^{K} \left(\beta_{kh}^{ij} \cdot \log [P(d_i) \cdot P(z_h|d_i) \cdot P(g_k|z_h) \cdot N(f_j|\mu_k, \Sigma_k)] \right) \quad (8)$$

Then, the update equations for the M-step become:

$$\mu_k^{new} = \frac{1}{p_k} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{h=1}^{H} \beta_{kh}^{ij} \cdot f_j \quad (9)$$

$$\Sigma_k^{new} = \left(\frac{1}{p_k} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{h=1}^{H} \beta_{kh}^{ij} \cdot f_j^2 \right) - (\mu_k^{new})^2 \quad (10)$$

[1]Derivations of the EM equations will be published simultaneously in a technical report with this paper.

where

$$p_k = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \sum_{h=1}^{H} \beta_{kh}^{ij} \quad (11)$$

and

$$P(z_h|d_i)^{new} = \frac{\sum_{j=1}^{N_i} \sum_{k=1}^{K} \beta_{kh}^{ij}}{N_i} \quad (12)$$

$$P(d_i)^{new} = \frac{N_i}{\sum_i N_i} \quad (13)$$

$$P(g_k|z_h)^{new} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} \beta_{kh}^{ij}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \beta_{kh}^{ij}} \quad (14)$$

In fact the solution to $P(d_i)$ is trivial, thus we will not have to estimate this distribution.

In order to estimate $P(z_h|d_i)$ for test images, we fix the learned Gaussian mixtures, i.e. $P(g_k|z_h)$ and the associated Gaussian distributions, i.e. Σ_k and μ_k, and perform the remaining steps of the above algorithm.

As the iterative EM-algorithm does not necessarily converge to the optimal solution, it is important to initialize the model, especially the parameters of the Gaussian distributions, appropriately in order to avoid local minimums. We initialize the means and covariances of the Gaussians representing the visual vocabulary by computing an ordinary multivariate Gaussian mixture model of the same size using all local features extracted in our training images. The topic and component probabilities are initialized randomly. It should be noted that we consider only the case of diagonal covariance matrices in our experiments.

4.2 FSGW-pLSA

In order to learn a FSGW-pLSA model, we perform exactly the same EM iteration steps as described in the previous subsection 4.1, but we do not update the μ_k's and Σ_k's of the Gaussian distributions in the M-step. Compared to the SGW-pLSA the FSGW-pLSA is less computational expensive, as the means and covariances of the Gaussian distributions do not have to be estimated in every the EM-step.

To derive the parameters of the Gaussians representing the fixed continuous vocabulary, we compute a multivariate Gaussian mixture model on the local feature vectors of the training set in advance. The Gaussian mixture model computation is initialized with the outcome of a k-means clustering on a feature subset of the training set. Note that again, we only consider the case of diagonal covariance matrices.

4.3 GM-pLSA

The log likelihood of the images in the database when using the GM-pLSA model is given by:

$$l = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \log \left(\sum_{h=1}^{H} \sum_{k=1}^{K} [P(z_h|d_i) \cdot P(d_i) \cdot \pi_{kh} \cdot N(f_j|\mu_{kh}, \Sigma_{kh})] \right) \quad (15)$$

As before, the existence of the sums inside the logarithm makes direct maximization of the log-likelihood by partial

derivatives difficult. Thus we again use the EM-algorithm to iteratively estimate the parameters. We derive the following update equation for the variables β_{kh} in the E-step:

$$\beta_{kh}^{ij} = \frac{P(z_h|d_i) \cdot \pi_{kh} \cdot N(f_j|\mu_{kh}, \Sigma_{kh})}{\sum_{h=1}^{H} \sum_{k=1}^{K} P(z_h|d_i) \cdot \pi_{kh} \cdot N(f_j|\mu_{kh}, \Sigma_{kh})} \quad (16)$$

The M-step updates result in:

$$\mu_{kh}^{new} = \frac{1}{p_{kh}} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \beta_{kh}^{ij} \cdot f_j \quad (17)$$

$$\Sigma_{kh}^{new} = \left(\frac{1}{p_{kh}} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \beta_{kh}^{ij} \cdot f_j^2 \right) - (\mu_{kh}^{new})^2 \quad (18)$$

where

$$p_{kh} = \sum_{i=1}^{M} \sum_{j=1}^{N_i} \beta_{kh}^{ij} \quad (19)$$

and

$$P(z_h|d_i)^{new} = \frac{\sum_{j=1}^{N_i} \sum_{k=1}^{K} \beta_{kh}^{ij}}{N_i} \quad (20)$$

$$P(d_i)^{new} = \frac{N_i}{\sum_i N_i} \quad (21)$$

$$\pi_{kh}^{new} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N_i} \beta_{kh}^{ij}}{\sum_{k=1}^{K} \sum_{i=1}^{M} \sum_{j=1}^{N_i} \beta_{kh}^{ij}} \quad (22)$$

Again, the solution to $P(d_i)$ is trivial and does not need to be estimated in our iterative algorithm. Computing $P(z_h|d_i)$ for test images is performed by keeping the parameters of the Gaussian mixtures fixed and only fitting the $P(z_h|d_i)$ parameters during the EM iterations.

An important aspect of this model is the choice of the number of Gaussian mixtures per topic. Here we compromise between the accuracy to represent the feature distributions per topic and the computational complexity as well as the ability to fit the model with a very large number of parameters. In addition, due to local maxima, special care has to be taken to initialize the model appropriately. In this work we initialize the parameters of the Gaussian mixtures by using the result of the SGW-pLSA. Only the K most important Gaussians per topic, i.e. the Gaussians with the highest probability of occurrence in each topic, are chosen. All other parameters are initialized randomly. Again we only consider the case of diagonal covariance matrices.

5. EXPERIMENTAL EVALUATION

5.1 Experimental Setup

We evaluate the three proposed continuous vocabulary models by means of scene recognition experiments on the often used OT dataset. The OT dataset [14] consists of a total of 2688 images from 8 different scene categories: coast, forest, highway, inside city, mountain, open country, street, and tall building. Table 1 and Figure 4 show the number of images as well as sample images for each category, respectively. For performance evaluation, each test image is assigned automatically by our system to one of the eight categories, and the achieved recognition rate is used as the performance measure throughout our experiments.

Figure 4: Example images per category of OT dataset.

category	scene type	# of images
1	coast	360
2	forest	328
3	highway	260
4	inside city	308
5	mountain	374
6	open country	410
7	street	292
8	tall building	356
total		2688

Table 1: List of the categories and their respective number of images in the OT dataset.

For evaluation we divide the images randomly into 1344 training and 1344 test images. We further subdivide the 1344 training images in a training and a validation set of size 1238 and 106, respectively. The validation set is used to find the best parameter configuration for the respective pLSA-based model. In the model we fix the number of topics to 25 and optimize the number, K, of visual words/Gaussian distributions as well as the number of EM iterations performed for the different models. It should be noted that pLSA-related models are susceptible to overfitting, thus an early termination may help with this issue. A number of 25 topics has been shown to result in good performance on this dataset [5].

Having determined for each model the best parameter setting for the number of visual words and EM iterations, we pick the according model and apply it to the entire training set (i.e., the set resulting from merging training and validation set). The model is also applied to all test set images in order to compute a topic distribution for each image. This topic vector for each image is then used to determine the most similar images in the training set to each query (test) image thus to finally determine the test image's category by the k-Nearest Neighbor algorithm.

Scene recognition is performed on all images in the test set. Based on these recognition results we compare the different

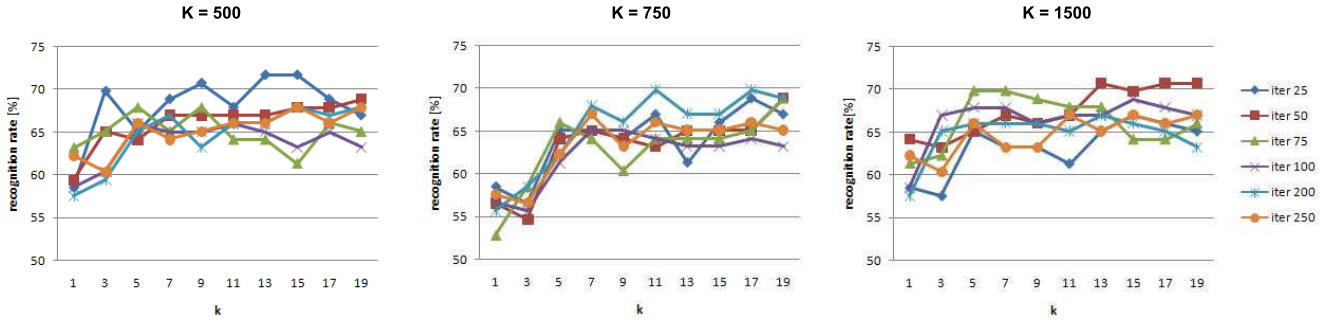


Figure 5: Recognition rates of the original pLSA model on the validation set for various k's of the kNN algorithm, different numbers of iterations of the EM algorithm, and different numbers of visual words K in the model.

proposed models. We use the performance of the discrete pLSA model as a baseline.

5.2 pLSA

The recognition rates for different numbers of visual words K, different k's of the kNN algorithm, and different numbers of EM iterations on the validation set are depicted in Figure 5. We can see that the best recognition rates on the validation set are achieved for a vocabulary size of 500 visual words and 25 iterations. The best recognition rate of approximately 71% is obtained for $k = 13$ and $k = 15$.

The results of the original pLSA model on the test set using the entire training set for $K = 500$ and 25 EM iterations will serve in Subsection 5.6 as a baseline for the evaluation of the proposed pLSA models with continuous vocabulary representations.

5.3 SGW-pLSA

Next we perform the above experiments with varying parameter configurations for the proposed SGW-pLSA model. The results are displayed in Figure 6. We clearly see that the results for a visual vocabulary size of $K = 1500$ are better than the ones obtained for 500 and 750 Gaussian distributions in the mixtures. A recognition rate of about 76% is achieved for $K = 1500$, $k = 15$ and 200 EM iterations. Thus we choose this parameter setting for computing the results on our test set in Subsection 5.6.

It can be also seen in Figure 6 that the model needs about 100 iterations to stabilize its performance. Thereafter the performance improves only slightly – in some cases even gets slightly worse. This could be a sign of overfitting. It should also be notes that a training set size of 1238 images, each producing in average about 550 local descriptors, is not very large for the number of parameters that must be estimated for a SGW-pLSA model containing 1500 Gaussian distributions.

5.4 FSGW-pLSA

Figure 7 shows the results obtained for the FSGW-pLSA model on the validation set. Again the size of the vocabulary, the parameter k in the kNN algorithm and the number of EM iterations in model estimation have been varied.

It can be seen that a vocabulary consisting of 750 and 1500 Gaussians give the best results. Especially the results for 1500 words and 250 iterations performed best with a recognition rate of about 72% for $k = 11$ and $k = 15$. As the results for $K = 1500$ and 250 iterations are close to the 70% for a larger range of k values compared to the results for $K = 750$ and 100 EM iterations, we will use the former parameter setting for the final model comparison on the test set.

5.5 GM-pLSA

In Figure 8 the recognition rates of the GM-pLSA on the validation set for various parameter settings are shown. The number of Gaussians K per topic ranges between 20 and 120. This results in a total number of between 500 and 3000 Gaussians in the model.

The results show that 20 and 30 Gaussians per mixture seem not to be sufficient as results improve with larger K. The best result is obtained for $K = 120$, i.e. a total number of 3000 Gaussian in the model, and 250 EM iterations. Here we obtain recognition rates of more than 72% for $k = 13, 15, 17$. Thus this parameter setting will be used to compute the performance of the model on the test set.

The results show that the total number of Gaussians in this model needs to be larger than in the previous examined models. Likely the results will further improve when going to even larger numbers of Gaussian distributions per topic. Nevertheless, this gets computational very expensive. The required larger total number of Gaussians might be explained by the fact that topics will still partly use similar Gaussians, but those have to be estimated for each topic separately.

5.6 Results

We will now compare the results of the different models using the parameter sets that have lead to the best performance on the validation set. We merge the training and validation set and use the computed models for these selected parameter sets to perform inference on the test set. Given the topic distribution on the test images, each test image is classified based on the dominant scene label in the k-Nearest Neighbor (kNN) set of the training images to the test image vector.

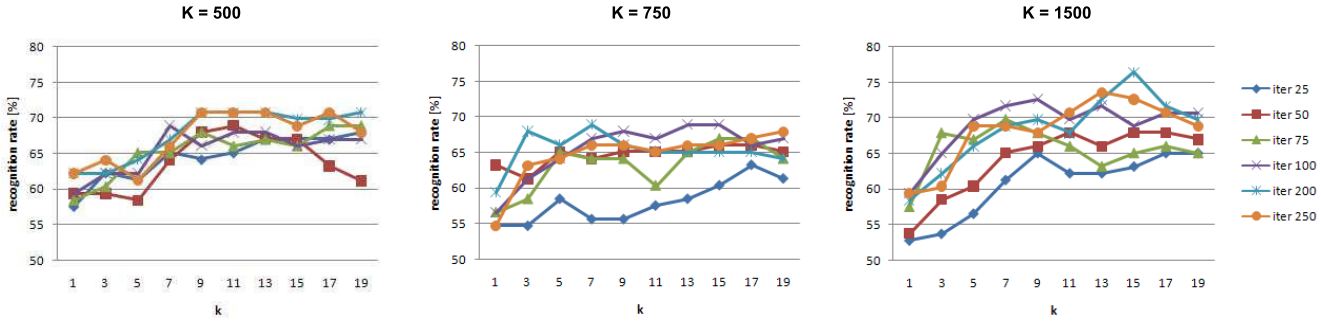


Figure 6: Recognition rates of the SGW-pLSA on the validation set for various k's of the kNN algorithm, different numbers of iterations of the EM algorithm, and different numbers of Gaussians K in the model.

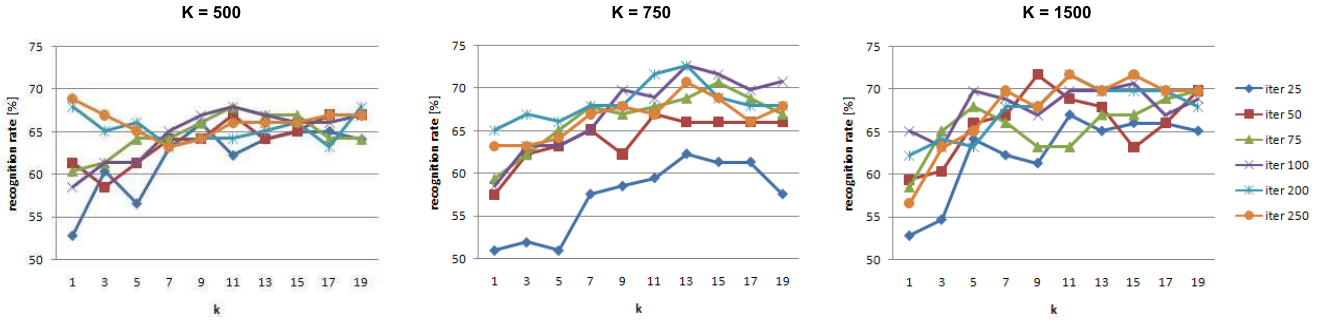


Figure 7: Recognition rates of the FSGW-pLSA on the validation set for various k's of the kNN algorithm, different numbers of iterations of the EM algorithm, and different numbers of Gaussians K in the model.

Figure 9 compares the achieved recognition rates on the test data set for different numbers of k of the nearest neighbor search. All three proposed continuous vocabulary models clearly outperform the original pLSA. The best performing model is the SGW-pLSA model, which only slightly outperforms the second best model, the FSGW-pLSA. Both approaches show a performance improvement of roughly 2% to 4% over the pLSA. The third best model, the GM-pLSA shows a recognition rate which is about 1% to 2% above the performance of the pLSA.

It should be noted that in the case of SGW-pLSA, we need to compute the parameters of 1500 Gaussians, whereas in the case of GM-pLSA we compute estimates for a total number of 3000 multivariate Gaussian distributions. Parameter optimization in Section 5.5 has shown that we do need this large number of Gaussians in the GM-pLSA to accurately model the database images. Nevertheless, the lower performance compared to the SGW-pLSA may be an result of having not enough training data to reliably learn this large number of parameters in the GM-pLSA model.

In summary, we conclude that a continuous pLSA model describes the visual environment better than a discrete pLSA model as used in previous topic model based scene recognition work. In this application domain the SGW-pLSA model and the FSGW-pLSA model outperform the GM-pLSA model, which has also the disadvantage of being computationally more expensive. Furthermore the performance improvement of the SGW-pLSA over the FSGW-pLSA is small, thus if low computational cost is required one should consider using the FSGW-pLSA over the SGW-pLSA.

6. CONCLUSION

In this paper we have proposed and evaluated three different extensions to the pLSA where the visual vocabulary is modeled by continuous feature vector distributions. For each of the models we have presented algorithms for parameter estimation and inference. A competitive evaluation in an automatic scene classification task shows that the proposed approaches outperform the discrete pLSA model. Furthermore, we found that the SGW-pLSA performed best closely followed by the FSGW-pLSA.
Future work will be to verify the results on a large scale dataset in an image retrieval task and using different local region detectors and descriptors in the experiments.

7. REFERENCES

[1] P. Ahrendt, C. Goutte, and J. Larsen. Co-occurrence models in music genre classification. In *IEEE International Workshop on Machine Learning for Signal Processing*, pages 247–252, 2005.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3, 2003.

[3] D. M. Blei and M. I. Jordan. Modeling annotated data. In *SIGIR '03: Proceedings of the 26th Annual*

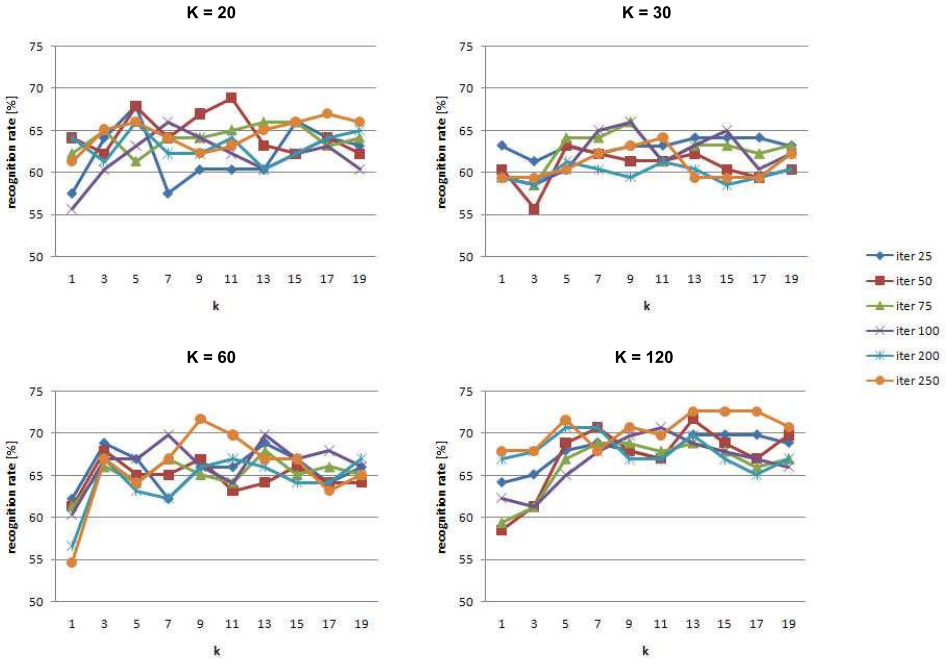


Figure 8: Recognition rates of the GW-pLSA on the validation set for various k's of the kNN algorithm, different numbers of iterations of the EM algorithm, and different numbers of Gaussians K in the model.

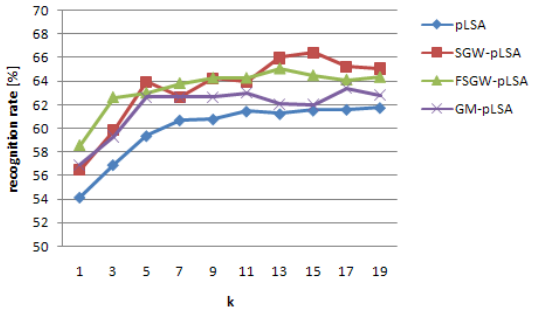


Figure 9: Recognition results for all models on the test set.

International ACM SIGIR conference on research and development in information retrieval, pages 127–134, 2003.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proceedings of the European Conference on Computer Vision*, 2006.

[6] L. Cao and L. Fei-Fei. Spatially coherent latent topic model for concurrent object segmentation and classification. In *IEEE Intern. Conf. on Computer Vision (ICCV)*, 2007.

[7] A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

[8] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, 42(1-2):177–196, 2001.

[9] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *CIVR '07: Proceedings of the 6th ACM international conference on image and video retrieval*, pages 17–24, 2007.

[10] D. Larlus and F. Jurie. Latent mixture vocabularies for object categorization. In *British Machine Vision Conference*, 2006.

[11] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 524–531, 2005.

[12] R. Lienhart and M. Slaney. pLSA on large scale image databases. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007.

[13] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[14] A. Oliva and A. B. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[15] 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 '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 883–890, 2005.

[16] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *International Conference on Computer Vision (ICCV 2005)*, 2005.

[17] M. Szummer and R. W. Picard. Indoor-outdoor image classification. In *IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, pages 42–51, 1998.

[18] A. Vailaya, M. Figueiredo, A. Jain, and H. Zhang. Image classification for content-based indexing. *IEEE Transactions on Image Processing*, 10(1):117–130, 2001.

[19] J. Vogel and B. Schiele. Natural scene retrieval based on a semantic modeling step. In *CIVR*, pages 207–215, 2004.

[20] S. Young. A review of large-vocabulary continuous-speech recognition. *IEEE Signal Processing Magazine*, 13(5):45–57, 1996.