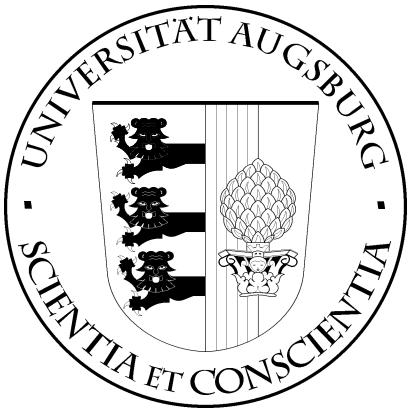
Fusing Local Image Descriptors for Large-Scale Image Retrieval

Eva Hörster, Rainer Lienhart

Angaben zur Veröffentlichung / Publication details:

Hörster, Eva, and Rainer Lienhart. 2007. "Fusing Local Image Descriptors for Large-Scale Image Retrieval." Augsburg: Universität Augsburg.

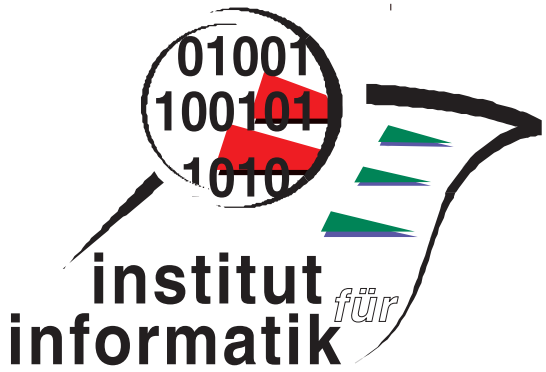
Universität Augsburg



Fusing Local Image Descriptors for Large-Scale Image Retrieval

E. Hörster, R. Lienhart

Report 2007-06 April 2007



Institut für Informatik
D-86135 Augsburg

Fusing Local Image Descriptors for Large-Scale Image Retrieval

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

Abstract

Online image repositories such as Flickr contain hundreds of millions of images and are growing quickly. Along with that the needs for supporting indexing, searching and browsing is becoming more and more pressing. Here we will employ the image content as a source of information to retrieve images and study the representation of images by topic models for content-based image retrieval. We focus on incorporating different types of visual descriptors into the topic modeling context. Three different fusion approaches are explored. The image representations for each fusion approach are learned in an unsupervised fashion, and each image is modeled as a mixture of topics/object parts depicted in the image. However, not all object classes will benefit from all visual descriptors. Therefore, we also investigate which visual descriptor (set) is most appropriate for each of the twelve classes under consideration. We evaluate the presented models on a real world image database consisting of more than 246,000 images.

1. Introduction

Nowadays there exist online image repositories containing hundreds of millions of images of all kinds of quality, size and content. One example of such an image repository is Flickr™. These image repositories grow day by day making techniques for navigating, indexing, and searching prudent. Currently indexing is mainly based on manually entered tags and/or individual and group usage patterns. Manually entered tags, however, are very subjective and not necessarily referring to the shown image content. This subjectivity and ambiguity of tags makes image retrieval based on manually entered tags difficult.

In this work we employ the image content as the source of information to retrieve images. It has been shown that recent probabilistic text models originally developed for large text document collections such as probabilistic Latent Semantic Analysis (pLSA) [7] and Latent Dirichlet Allocation (LDA) [3] improve retrieval performance in an image similarity search tasks on large real world databases [6, 9]. Previously those models were successfully applied and extended to image content analysis tasks such as scene classification [4, 8, 11], object categorization [5, 13, 14] and the problem of modeling annotated image collections [1, 2].

The above mentioned probabilistic text models describe documents as mixtures of intermediate hidden topics (also called aspects) under the assumption of a bag-of-words document representation. Given unlabeled training documents, the probability distributions of these models are estimated in a completely unsupervised fashion. In the visual domain the mixture of hidden topics refers to the degree to which a certain object/scene type is contained in the image and models therefore the co-occurrence of so called visual words inside and across images. In the ideal case, the mixture of topics in a specific image gives rise to a low-dimensional description of the coarse image content and thus enables retrieval in very large databases. It allows us to put images into subspaces for higher-level reasoning which in turn can be used to find similar images.

Building visual words using texture features that describe local image regions has been shown to work well in the image retrieval task [6, 9], but we believe that the results can be improved for some object categories and scenes types. Those categories are best modeled by fusing texture descriptors and a second type of visual feature. In this work we will consider color patches as the second type, but the approach works similar for other types. In the context of topic models various types of basic local image descriptors/visual words such as gray-scale patches, color patches or SIFT features have been investigated, but previous works have either considered only one local image description type in their models [8, 4] or different local image descriptors have been fused during feature generation on the feature level [1, 2]. In [12] Quelhas and Odobez study two fusion approaches to combine color and texture information in a bag-of-visual words representation.

In this work we will propose three approaches for fusion of different feature types in the context of topic models. Fusion will be carried out at different stages of the models: at the visual word level, during topic generation and at the decision level. We will evaluate the models experimentally by user studies in a retrieval-by-example task on a large-scale real world image database consisting of more than 246,000 images downloaded from the public Flickr repository.
As not all object classes benefit from visual descriptor fusion, we examine in a second step which categories are best modeled by only one feature type and which category models are improved by taking into account two different kinds of visual words. Here we will build on the insights gained in the first experiments with respect to the best fusion model.
We can summarize the main contributions of this paper as follows:

- We explore three different topic models for feature fusion and their application to content based image retrieval.
- We judge the suitability of the presented models by user studies on a real world, large-scale image database with more than 246,000 images.
- We examine different local descriptors and their combination with respect to their suitability to model certain image categories.

The paper is organized as follows. First, visual word computation for local image features is discussed in Section 2. Section 3 reviews the LDA model and introduces the proposed fusion models. Section 4 describes our dataset and introduces the similarity measure for finding images of similar content. The evaluation methodology is outlined and experimental results are given and discussed in Section 5. Section 6 concludes the paper.

2. Visual Word Computation

The first step in building a generative probabilistic model for our image collection is deriving the bag-of-visual words image representations. Therefore we need to compute a visual vocabulary consisting of N visual words for each local descriptor type.
We will now describe how a vocabulary for one type of local image descriptor is computed and which descriptor types are considered in this work. We will use the term *feature* and *descriptor* interchangeably. The discussion on the different possibilities of fusing these feature types is postponed till Section 3.
A vocabulary is usually derived in two steps. First features are computed at predefined locations and scales. Then the vocabulary is built by vector quantizing the automatically extracted local image descriptors.

In this work we will consider two different possibilities of defining interest points and scales for feature extraction:

- *Sparse features*: Interest points are detected at local extremas in the difference of Gaussian pyramid [10]. A position and scale are automatically assigned to each point and thus the extracted regions are invariant to these properties.
- *Dense features*: Interest points are defined at evenly sampled grid points. Feature vectors are then computed based on three different neighborhood sizes, i.e. at different scales, around each interest point. These three different scales should allow for a (very) limited degree of scale invariance in the representation.

Two kinds of visual features are computed for describing a detected region of interest: color patch features and rotation invariant SIFT features. Color patch features are computed from normalized 7×7 pixels RGB patches. For each color channel a 49-dimensional feature vector is computed from the patches' pixel values. By combining the values from all three channels we obtain a 147-dimensional features vector. The well-known SIFT features [10] are computed by first assigning an orientation to each interest point. Then we compute a 128-dimensional gradient-based feature vector from the local grayscale neighborhood of each interest point in an orientation invariant manner.
In previous work [9] the authors investigated three techniques for learning visual words from local image features for large-scale image databases. We use the best performing technique in this work for visual word computation: merging the results of multiple k-means clustering on non-overlapping feature subsets. Therefore relatively small sets of features (compared to the entire number of features in all 246,000 images) are selected randomly from all features. Then k-means clustering is applied to each subset and the means of each cluster are kept as visual words. Finally, the derived visual words of each subset are amalgamated into the vocabulary. This approach is several magnitudes computationally more efficient compared to determining all clusters from one large set of features.
Given the vocabulary for each feature type, we describe each image as a collection of visual words by replacing each detected feature vector in the respective image by its most similar visual word of the same type: the most similar is defined as the closest word in the 128-dimensional (SIFT) or 149-dimensional (color patch) vector space. Since the order of terms in a document is ignored, any geometric relationship between the occurrences of visual words in images is disregarded. Such a model is widely known as a bag-of-(visual) words model.
Our aim in this work is to investigate the possibilities of fusing different types of visual words in the context of topic models. We will limit our studies to the case that in each

image I_d the same number of N_d (depending on the images' size, texture, etc.) color patch and SIFT features are extracted. Moreover color patch and SIFT features fused in our models are extracted at the same interest points and with the same scale. Thus we will consider color patch and SIFT words either both densely detected or both sparsely detected. This procedure enables us to fuse image descriptors directly at the word level where color patch and SIFT word occurrence at the same interest point are directly fused while building the bag-of-words model (see fusion model B, Section 3.3).

3. LDA-based Fusion Models

As stated in the introduction, there exist two probabilistic text models that use hidden topics to model document collection: the pLSA [7] and the closely related LDA [3]. Compared to pLSA, the LDA provides a completely generative model and therefore overcomes some problems of the pLSA. Moreover LDA has been shown in [6] to perform superior to pLSA in a content-based image retrieval task on a large-scale database. Thus we will build our feature fusion approaches on this model. Nevertheless, the proposed models can be applied analogously to the pLSA. Before outlining our feature fusion models we will first review the original LDA model.

3.1. LDA Model

Latent Dirichlet Allocation (LDA) [3] is a generative probabilistic model developed for collections of text documents. It represents documents by a finite mixture over latent topics, also called hidden aspects. Each topic in turn is characterized by a distribution over words and each occurrence of a word in a specific document is associated with one unobservable topic. In this work our aim is to model image databases not text databases, thus our documents are images and topics correspond to objects depicted in the images. Most importantly LDA allows us to represent an image as a mixture of topics, i.e. as a mixture of multiple objects.

In order to apply the original LDA model to image databases, each image I_d is represented by a bag-of-words model, i.e. as a sequence of N_d visual words w_n, written as $\mathbf{w_d} = \{w_1, w_2, ..., w_{N_d}\}$. Then the process of generating such an image is described as follows [3]:

- Choose a K-dimensional Dirichlet random variable $\theta \sim Dir(\alpha)$, where K denotes the finite number of topics in the corpus.
- For each of the N_d words w_n:
  - Choose a topic $z_n \sim Multinomial(\theta)$
  - Choose a word w_n from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic z_n

The likelihood of an image I_d according to this model is given by:

$$p(\mathbf{w_d}|\alpha, \beta) = \int p(\theta|\alpha) \prod_{n=1}^{N_d} (\sum_{j=1}^{K} p(z_j|\theta) p(w_n|z_j, \beta)) d\theta \quad (1)$$

The probability of the complete image database is the product of the likelihoods of single images.

Probability distributions of visual words given a hidden topic as well as probability distributions of hidden topics given the images are learned in a complete unsupervised manner. We learn an LDA model by finding the corpus parameters α and β such that the log marginal likelihood of a database consisting of a number of training images is maximized. Since Eqn. 1 cannot be solved directly, model parameters are estimated by variational inference [3]. Given the learned parameters α and β, we assign probabilities to an image by maximizing the respective log marginal likelihood. Thus we may learn the LDA corpus level parameters on a subset of the database (in order to reduce total training time) and then assign probability distributions to all images.

3.2. Fusion Model A

Our first proposed fusion model consists basically of two completely independent learned LDA representations for the images in the database. One LDA model is learned for the bag-of-words image representation based on the color patch vocabulary and one for the representation based on SIFT features. The fusion is performed at the decision level, i.e. topic distributions are computed independently and fusion of those two LDA models is carried out while measuring similarity during retrieval (see Section 4.2).

It should be noted that in this model topics are not 'shared' between features. Thus a topic is either purely a color patch topic or a topic defining a distribution over texture words. Topics, which are characterized by both color and texture, are not properly modeled here. However, the separation might be beneficial if combined with some active learning retrieval system. Such as system could learn whether one or both features and thus the corresponding topics are important to find images of similar content.

The graphical representation of the LDA-based fusion model A is shown in Figure 1(a). M indicates the number of images in the entire database and N_d denotes the number of visual words of each feature type that are detected in image I_d.

3.3. Fusion Model B

The second model fuses the feature types at the visual word level and assumes a joint observation of a color patch word and a SIFT word. Thus, each time a topic z_n is chosen, a color-patch word c_n and a SIFT word t_n – both coming

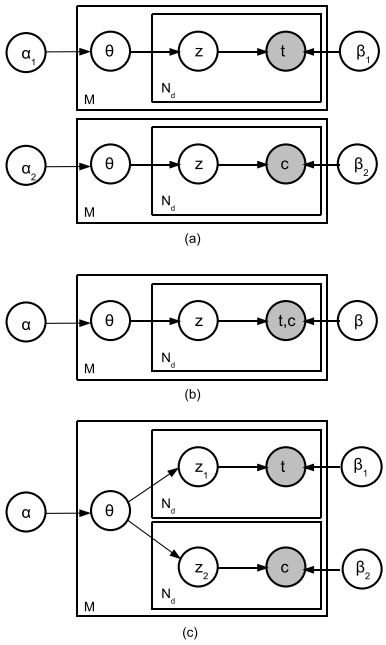


Figure 1. Graphical representation of the LDA-based fusion models: (a) fusion model A; (b) fusion model B; (c) fusion model C. (M denotes the number of images in the database and N_d the number of detected visual words of a certain feature type in image I_d. The shaded nodes denote the observable random variables c and t for the occurrence of a color patch or SIFT word, respectively. z denotes the topic variable and θ the topic mixture variable.)

from the same interest point and scale – are sampled from a multinomial probability conditioned on the topic z_n. Here we explore the fact that in each image we compute color patch features and SIFT feature at the same locations and scales, resulting in the same number of features for both types.

In this model we have a joint distribution over color and texture words for each topic. The likelihood of the occurrence of a combination of a specific texture word t_n and a color patch word c_n in an image according to this model is then given by:

$$p(t_n, c_n|\alpha, \beta) = \int p(\theta|\alpha)[\sum_{j=1}^{K} p(z_j|\theta) \cdot p(t_n, c_n|z_j, \beta)]d\theta \quad (2)$$

Note that this model does not allow topics representing only visual words of one feature type, as visual words are already fused at the word level.

The graphical representation of the LDA-based fusion model B is shown in Figure 1(b).

3.4. Fusion Model C

The third model aims to enable topics to represent either words of only one of the feature types or a combination. Here the latent topics for each sampled visual word (either color-patch or SIFT) can vary while the topic mixture θ is fixed, thus θ denotes a probability distribution over the hidden topics and in turn each visual word is originated from one of those topics. This is nothing else than concatenating the collection of visual words of both types to describe an image I_d, i.e. we can represent I_d by $\mathbf{w_d} = \{t_1, t_2, ..., t_{N_d}, c_1, c_2, ..., c_{N_d}\}$. The likelihood of an image I_d according to this model is then given by:

$$p(\mathbf{w_d}|\alpha, \beta) = \int p(\theta|\alpha) \left(\prod_{n=1}^{N_d} \sum_{j=1}^{K} p(z_{1j}|\theta) p(t_n|z_{1j}, \beta_1) \right) \cdot \left(\prod_{n=1}^{N_d} \sum_{j=1}^{K} p(z_{2j}|\theta) p(c_n|z_{2j}, \beta_2) \right) d\theta \quad (3)$$

The graphical representation of the LDA-based fusion model C is shown in Figure 1(c).

It should be noted that although the model allows topics purely representing words of one type of local descriptor, describing images that contain objects only characterized by one feature type (e.g. texture) is not possible as every visual word needs be 'explained' by one topic. Thus the topic distribution will have to account for words based on the second visual descriptor type (e.g. color patch words), too. This problem could be solved by using a relevance feedback algorithm.

Parameters of all three fusion models are calculated by variational inference as described in [3]. Again, learning the models involves finding the parameters α_i and β_i such that the log marginal likelihood of the training set is maximized. Probabilities are assigned to all images in the database by maximizing the log marginal likelihood of the respective image given the corpus level parameters.

4. Database and Similarity Measure

The objective of example-based image retrieval is to obtain images with content similar to a given query image. We evaluate results purely based on the visual similarity of the retrieved images as perceived by ordinary users.

4.1. Database

All experiments are performed on a database consisting of approximately 246,000 images. The images were selected from all public Flickr images uploaded prior to Sep. 2006 and labeled as *geotagged* together with one of the following tags: *sanfrancisco, beach* and *tokyo*. Of these im-

Category	*OR* list of Tags	# of images
1	wildlife animal(s) cat(s)	28509
2	dog(s)	24660
3	bird(s)	20908
4	flower(s)	25457
5	graffiti	21888
6	sign(s)	14333
7	surf(ing)	29552
8	night	33142
9	food	18602
10	building(s)	16826
11	goldengate goldengatebridge	23803
12	baseball	12372
	Total # of images (Note images may have multiple tags)	246,348

Figure 2. Image database and its categories used for experiments

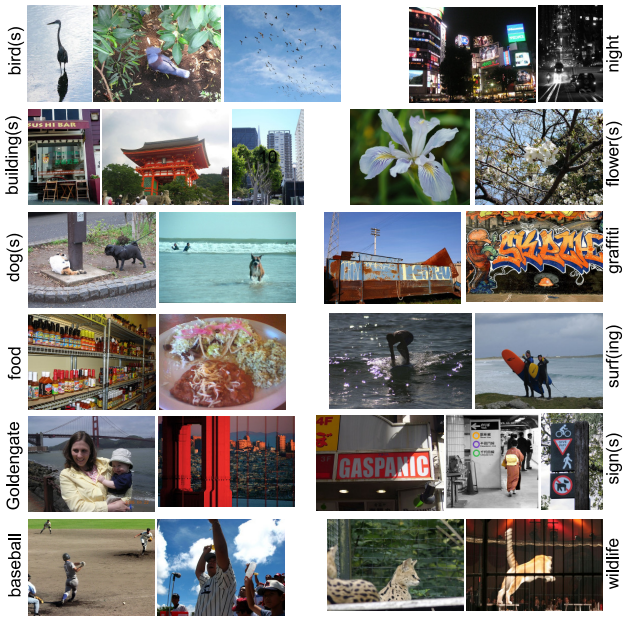


Figure 3. Example images from the 12 different categories of the Flickr dataset

ages only images having at least one of the following tags were kept: *wildlife, animal, animals, cat, cats, dog, dogs, bird, birds, flower, flowers, graffiti, sign, signs, surf, surfing, night, food, building, buildings, goldengate, goldengatebridge, baseball*. The resulting image database was not cleaned nor preprocessed in any way to increase consistency and thus we can group images into 12 categories as shown in Figure 2. Example images from all 12 categories are shown in Figure 3.

The preselection of a subset of images from the entire Flickr database based on tags is needed as Flickr is a repository with hundreds of millions of images. However, it should be noted, that indexing purely based on tags is not sufficient as the tags are a very noisy indication of the content shown in the images. This can be observed in Figure 4.

Note that this database has been also used for the experimental evaluation in [6, 9].

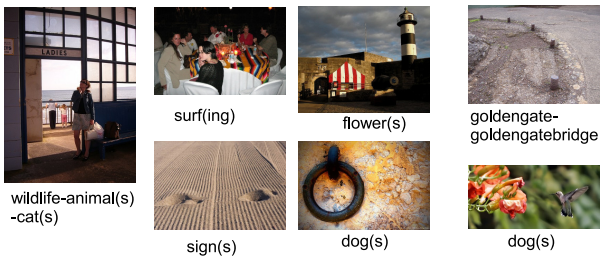


Figure 4. Example images from different categories: categories/tags do not refer to the content shown

4.2. Image Similarity Measure

We focus on the task of query-by-example, thus searching in the database for the most similar items to a given query image. Once we have trained an LDA model or one of the LDA-based fusion models and computed a probabilistic representation for each image in the database based on those, we need to define a similarity measures in order to perform image retrieval.

The topic mixture θ indicates to what degree a certain topic is contained in the respective image. In previous work [6] various similarity measures for image retrieval based on topic mixtures have been investigated and we adopt in this work the measure that has been shown to be the best performing measure. This measure has been adopted from language based information retrieval. Each document is indexed by the likelihood of its model generating the query document, i.e. the most relevant documents are the ones whose model maximizes the conditional probability on the query terms. In content-based image retrieval, a query image I_a can be presented as a sequence of visual words $\mathbf{w_a}$ and thus the above mentioned likelihood can be written as:

$$P(\mathbf{w_a}|M_b) = \prod_{i=1}^{N_d} P(w_i^a|M_b) \tag{4}$$

where M_b is the model of an image I_b and N_d the total number of detected visual words in image I_a.

Applying this measure to our three fusion model it turns into:

Fusion Model A:

$$P(\mathbf{w_a}|M_b) = \prod_{i=1}^{N_d} P(t_i^a|M_b^t) \cdot \prod_{i=1}^{N_d} P(c_i^a|M_b^c) \tag{5}$$

We have computed two independent LDA-models for each type of visual vocabulary, thus we have two models for image I_b, M_b^t denotes the model based on the texture vocabulary and M_b^c the one stemming from the color patch vocabulary, respectively. The total number of visual words in one image is given by $2 \cdot N_d$ as we extract N_d color patches and the same number of SIFT features in image I_a.

Fusion Model B:

$$P(\mathbf{w_a}|M_b) = \prod_{i=1}^{N_d} P(t_i^a, c_i^a|M_b) \quad (6)$$

Here each term w_i^a in the document is build from a combination of a color patch and a SIFT word, i.e. $w_i^a = \{t_i^a, c_i^a\}$. Each image gives rise to N_d combined terms.

Fusion Model C:

$$P(\mathbf{w_a}|M_b) = \prod_{i=1}^{N_d} P(t_i^a|M_b) \cdot \prod_{i=1}^{N_d} P(c_i^a|M_b) \quad (7)$$

Again each image I_d is represented as a collection of $2 \cdot N_d$ visual words. Compared to model A, we have also two kinds of words, but only one model.

Wei and Croft [15] combine the LDA model and a simple unigram model with Dirichlet smoothing to estimate the terms $P(w_i^a|M_b)$ in order to perform information retrieval. We will now combine the LDA-based fusion models instead of the LDA model with this measure:

$$P(w_i^a|M_b) = \lambda \cdot P_u(w_i^a|M_b^u) + (1-\lambda) \cdot P_{f_m}(w_i^a|M_b^{f_m}) \quad (8)$$

where $P_u(w_i^a|M_b^u)$ is specified by the unigram document model with Dirichlet smoothing according to [16]:

$$P_u(w_i^a|M_b^u) = \frac{N_d^b}{N_d^b + \mu} P_{ML}(w_i^a|M_b^u) + (1 - \frac{N_d^b}{N_d^b + \mu}) P_{ML}(w_i^a|D) \quad (9)$$

D denotes the entire set of images in the database, μ the Dirichlet prior and N_d^b the number of visual words in image I_b. The maximum likelihood probabilities $P_{ML}(w_i^a|M_b^u)$ and $P_{ML}(w_i^a|D)$ are measured separately for each vocabulary type if model A or model C is considered. For model B those likelihoods are calculated for the joint visual words $\{t_i^a, c_i^a\}$.

The term $P_{f_m}(w_i^a|M_b^{f_m})$ in Eq. 8 refers to the probability of a visual word (combination) w_i^a in image I_a given the currently considered fusion model $M_b^{f_m}$ of image I_b. These probabilities are given by:

Fusion Model A:

$$P_{f_A}(w_i^a|M_b^{f_A}) = \sum_{j=1}^{K} P(w_i^a|z_j, \beta) \cdot P(z_j|\theta^b, \alpha) \quad (10)$$

where w_i^a may denote a color c_i^a or texture t_i^a word and the according LDA model representation of image I_b, i.e. its topic mixture θ^b, is applied.

Fusion Model B:

$$P_{f_B}(w_i^a|M_b^{f_B}) = P_{f_B}(c_i^a, t_i^a|\alpha, \theta^b, \beta) = \sum_{j=1}^{K} P(c_i^a, t_i^a|z_j, \beta) \cdot P(z_j|\theta^b, \alpha) \quad (11)$$

Fusion Model C:

$$P_{f_C}(w_i^a|M_b^{f_C}) = \sum_{j=1}^{K} P(w_i^a|z_j, \beta) \cdot P(z_j|\theta^b, \alpha) \quad (12)$$

where w_i^a denotes either a color word c_i^a or a texture word t_i^a and the corresponding β has to be inserted.

5. Experimental Results

For both feature types we computed a visual vocabulary from 12 randomly selected non-overlapping subsets each consisting of 500,000 local features. Each of those subsets produces 200 visual words giving a total vocabulary size of 2400 visual words for each type. In order to keep the overall number of visual words approximately constant, we compute for fusion model B only 70 visual SIFT words and 70 color patch words, giving in total 4900 possible combinations of SIFT and color patch words. Vocabularies are computed for sparsely and densely extracted features separately.

The LDA-based fusion models are learned on a training corpus consisting of 25,000 randomly chosen images from the dataset. The number of topics was set to 100 in fusion model B and C, whereas it was chosen to 50 in each of the two LDA models in fusion model A. This also gives in total 100 topics, 50 for the color-patch based model and 50 for the SIFT based model.

The Dirichlet prior μ in Equation 9 was set to 50 for our experiments.

5.1. Evaluation Methodology

We judge the performance of the different fusion models by users in a query-by-example task: We selected five query images per category at random resulting in a total of 60 query images. For each query image the 19 most similar images derived by the distance measure presented in the previous section are presented to the users. The users were asked to judge the retrieval results by counting how many of the retrieved images show content similar to the query image. As the query image is counted too, the lowest number of correctly retrieved images will be one and the largest 20. The average number of similar images over all categories is computed for each user to give the final result.

In the second part of our experimental evaluation we will study different local descriptors and their combination with

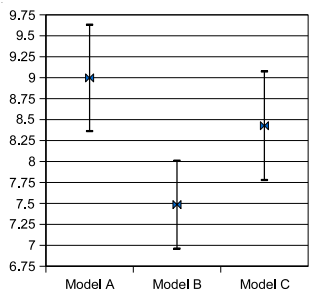


Figure 5. Resulting scores for the comparsion between the three fusion models applied to sparsely extracted features

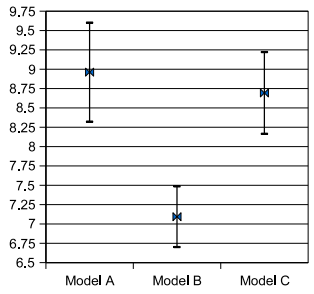


Figure 6. Resulting scores for the comparsion between the three fusion models applied to densely extracted features

Figure 7. Retrieval results obtained by our fusion models. The left most image in each row shows the query image; the four images to the right show the four most similar images.

respect to their suitability to model various image categories. In these experiments we selected 10 images randomly per category from our database (see Figure 2) [1] and compared the retrieval results obtained by the best performing fusion model to the retrieved images by an LDA image representation based on color patch features and one based on SIFT features. For this purpose we trained two 50 topic LDA models, one on the SIFT bag-of-words representation and another on the color patch representation. The distance measure in Section 4.2 was modified appropriately (for details see [6]). Evaluation is again performed by user studies as described above, except that the average is computed per category as each category is treated separately.

5.2. Fusion Models

In this section our aim is to evaluate the proposed fusion models. We performed two experiments: In the first one we compared the retrieval results obtained by the models using sparse features as the basic building block, while in the second experiment the models obtained from densely extracted features were used. The results of both experiments are depicted in Figure 5 and Figure 6. The vertical bars mark the standard deviation of the eight test users' scores.
In both experiments model A performs best followed by model C. Model B shows the worst performance. The results indicate that computing two separate LDA-models for image representation – one for each feature type – and fusing the information at the decision level (late fusion) gives the best results even in the unsupervised retrieval task. Moreover, the computational complexity is lower for model A.
Figure 7 displays some retrieval results obtained with the proposed models. The top six rows depict examples where the system works very well. The following three lines are examples where the returned retrieval results can be improved.

[1] The randomly obtained images were filtered to fit the category. For instance a cat image in the category buildings was skipped. Why a cat image was tagged "building" remains mysterious to the authors.

5.3. Model Selection

Having determined the most appropriate fusion approach, we will now examine the two different local descriptors, color-patches and SIFT, as well as their combination with respect to their suitability to model certain image categories. Therefore we consider the twelve categories in our database separately. Figure 8 and Figure 9 show the results for sparse and dense feature extraction, respectively. The average scores over five test users are depicted and the most suitable model is marked in yellow.
As expected, categories that are highly textured such as *graffiti* and *signs* are best modeled by a SIFT-based LDA model. The *wildlife* category contains also many textured objects such as tigers and lions, whereas the *bird* category is best described by color and shape and thus benefits from the

category	color-patch	model A	SIFT
wildlife animal(s) cat(s)	2.04	3.56	3.68
dog(s)	3.68	5.24	5.16
bird(s)	4.46	5.08	3.90
flower(s)	11.30	11.78	6.40
graffiti	4.06	7.04	10.38
sign(s)	2.98	3.52	3.86
surf(ing)	8.44	11.28	8.24
night	1.86	3.74	3.74
food	5.18	6.70	3.64
building(s)	2.32	2.56	2.56
goldengate(bridge)	4.38	8.08	10.96
baseball	15.82	16.56	12.32

Figure 8. Average scores per category for the comparsion between retrieval results based on LDA (fusion) models applied to sparsely extracted features

category	color-patch	model A	SIFT
wildlife animal(s) cat(s)	2.26	4.08	4.16
dog(s)	4.44	4.78	5.16
bird(s)	5.44	5.86	4.52
flower(s)	9.10	9.82	4.68
graffiti	4.52	5.92	7.50
sign(s)	2.38	2.98	5.12
surf(ing)	8.00	11.20	7.62
night	3.46	3.94	3.90
food	5.54	5.10	4.14
building(s)	3.22	2.90	2.68
goldengate(bridge)	7.88	9.02	8.68
baseball	14.16	14.80	16.66

Figure 9. Average scores per category for the comparsion between retrieval results based on LDA (fusion) models applied to densely extracted features

fusion of color patches (which model color as well as intensity changes) and SIFT features. *Flower* retrieval is also improved by the fusion. Altogether, the resulting scores show that many categories benefit from the fusion of both models.

Color patches alone are not appropriate for category modeling, as they only show superior performance in the two categories *food* and *building(s)* if dense feature extraction is considered. It should be noted that the standard deviation between users were large in the *building(s)* and in the *sign(s)* category indicating that the shown content was not obvious and thus it was diversely interpreted by the test users.

6. Conclusions

In this work we studied the fusion of two feature types in the context of topic models for query-by-example image retrieval. The three proposed approaches fuse the features at the visual word level, at the topic level or at the decision level. A probabilistic similarity measure was adopted and two feature detection methods were considered separately: dense and sparse detection. The experimental evaluation has shown that the fusion at the decision level performs best. Furthermore, the experiments show that some categories benefit from the fusion of local descriptor types while other categories are better modeled by only one feature type. Future work will include the verification of the results by a larger amount of users, categories and images per category.

References

[1] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[2] D. M. Blei and M. I. Jordan. Modeling annotated data. In *Proc. SIGIR*, pages 127–134, New York, NY, USA, 2003. ACM Press.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.

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

[5] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning object categories from google's image search. In *Proc. ICCV*, pages 1816–1823, Washington, DC, USA, 2005. IEEE Computer Society.

[6] E. Hörster, R. Lienhart, and M. Slaney. Image retrieval on large-scale image databases. In *ACM CIVR*, 2007.

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

[8] F.-F. Li and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proc. CVPR*, pages 524–531, Washington, DC, USA, 2005. IEEE Computer Society.

[9] R. Lienhart and M. Slaney. pLSA on large scale image databases. In *ICASSP*, 2007.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[11] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, and L. V. Gool. Modeling scenes with local descriptors and latent aspects. In *Proc. ICCV*, pages 883–890, Washington, DC, USA, 2005. IEEE Computer Society.

[12] P. Quelhas and J.-M. Odobez. Natural scene image modeling using color and texture visterms. In *Proc. CIVR*, pages 411–421, 2006.

[13] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *Proc. ICCV*, 2005.

[14] G. Wang, Y. Zhang, and L. Fei-Fei. Using dependent regions for object categorization in a generative framework. In *Proc. CVPR*, pages 1597–1604, Washington, DC, USA, 2006. IEEE Computer Society.

[15] X. Wei and W. B. Croft. LDA-based document models for ad-hoc retrieval. In *Proc. SIGIR*, pages 178–185, New York, NY, USA, 2006. ACM Press.

[16] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proc. SIGIR*, pages 334–342, New York, NY, USA, 2001. ACM Press.