Comparing Local Feature Descriptors in pLSA-Based Image Models

Hörster[1], Thomas Greif[1], Rainer Lienhart[1], and Malcolm Slaney[2]

[1] Multimedia Computing Lab, University of Augsburg, Germany
{hoerster,lienhart}@informatik.uni-augsburg.de
[2] Yahoo! Research, Santa Clara, CA, USA
malcolm@ieee.org

ostract. Probabilistic models with hidden variables such as proba-istic Latent Semantic Analysis (pLSA) and Latent Dirichlet Allocation DA) have recently become popular for solving several image content alysis tasks. In this work we will use a pLSA model to represent images performing scene classification. We evaluate the influence of the type local feature descriptor in this context and compare three different criptors. Moreover we also examine three different local interest re-n detectors with respect to their suitability for this task. Our results w that two examined local descriptors, the geometric blur and the f-similarity feature, outperform the commonly used SIFT descriptor a large margin.

roduction

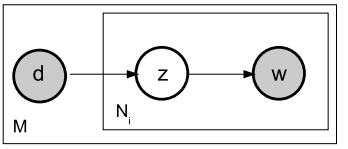
istic models with hidden topic variables, originally developed for sta-xt modeling of large document collections such as probabilistic Latent Analysis (pLSA) [1] and Latent Dirichlet Allocation (LDA) [2], have become popular in image content analysis tasks such as scene classi-3,4,5], object recognition [6], automatic segmentation [7] and image on [8]. In these approaches documents are modeled as mixtures of hid-cs under the assumption of a bag-of-words document representation. to visual tasks, the mixture of hidden topics refers to the degree to ch object class, i.e. grass, people, sky, is contained in the image. In the e, this gives rise to a low-dimensional image description of the coarse ntent, making the description particularly suitable for tasks such as rieval [9,10] and scene classification [3,4,5]. Hidden topic model based presentations outperform in both tasks previous approaches [9,4].

applying topic models in the image domain, the first step is to find an ate visual equivalent for words in documents. This is usually done by g local images descriptors computed for each image. A wide variety of ocal descriptors has been proposed [11,12,13,14] and they have become

n between advanced local descriptors in the context of pLSA models
issing. In a matching task, the aim is to find precisely corresponding
an object or scene in two images under different viewing conditions
ghtning or pose changes. This requires a very distinct region descrip-
vever, in a pLSA based scene classification or image retrieval task we
e to pool features describing visually similar regions in order to produce
ul visual words. Previous works on pLSA based image models only ap-
compared the popular SIFT [11] descriptor or simple color/gray scale
3,4,5]. Bosch et al.'s work [4] proposes a variation of SIFT, taking color
into account, in the context of scene recognition with a pLSA model
age representation.
s work we compare two recently proposed local features descriptors,
etric blur descriptor [13] and the self-similarity descriptor [14] in a
ssification task using a pLSA-based image representation. Both features
wn promising performance in image analysis tasks and have not been
d in the previous comparison [15]. As the SIFT based descriptors have
outperform other features [15], we take results obtained with the SIFT
r as a baseline and we use the classification rate on a previously unseen
as a performance measure. Moreover we also evaluate three different
rest region detectors with respect to their suitability for this task.

proach

ork we use a pLSA model to represent each image [4]. pLSA [1] was
derived in the context of text modeling, where words are the elemen-
s of documents. The starting point for building a pLSA model is to first
the entire corpus of documents by a term-document co-occurrence ta-
e $M \times N$. M indicates the number of documents in the corpus and N
ber of different words occurring across the corpus. Each matrix entry
e number of times a specific word (column index) is observed in a given
t (row index). Such a representation ignores the order of words/terms
ment and is commonly called a *bag-of-words* model.
er to be able to apply this model in the image domain, we first need to
isual equivalent to words in documents. Visual words are often derived
quantizing automatically extracted local region descriptors. This work
eans clustering on a subset of local features extracted from training
nd the cluster centers become our visual vocabulary.
the vocabulary, we extract local features from each image in the database
ce each detected feature vector with its most similar visual word, defined
sest word in the high-dimensional feature space. The word occurrences
ed, resulting in a term-frequency vector for each image document. These
uency vectors for each image then constitute the co-occurrence matrix.
order of terms in a document is ignored, any geometric relationship be-



raphical representation of pLSA model: $M = \#$ of images in database, $N_i = \#$ vords in image d_i, observable random variable (shaded) w for the occurrence l word and d for the respective image document, $z =$ hidden topic variable

the co-occurrence matrix, the pLSA uses a finite number of hidden model the co-occurrence of visual words inside and across images. ge is explained as a mixture of hidden topics and these hidden topics bjects or object parts. Thus we model an image as consisting of one le objects: e.g., an image of a beach scene consists of water, sand and ssuming that every word w_j occurring in a document d_i in the corpus ated with a hidden, unobservable topic variable z_k, we describe the ty of seeing word w_j in document d_i by the following model:

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

$d_i)$ is the prior probability of picking document d_i and $P(z_k|d_i)$ the ty of selecting a hidden topic depending on the current document, also o as the topic vector. Figure 1 shows the graphical representation of model.

rn the probability distributions of visual words given a hidden topic, the probability distributions of hidden topics given a document, com- nsupervised using the Expectation Maximization (EM) algorithm [1,16]. ty distributions of new images that are not contained in the original corpus are estimated by a fold-in technique [1]. Here the EM algorithm l to the unseen images to compute its topic distribution while keeping distributions conditioned on the topic $P(w_j|z_k)$ fixed. In our work, we the parameters of a pLSA model on the training data and then apply el to the test data using the fold-in technique. Finally, we represent each its associated topic vector $P(z|d)$ which gives us a very low-dimensional presentation.

ene recognition the topic vectors of each unlabeled test image are clas- simple k-Nearest Neighbor (kNN) search through the labeled training sing the L2-norm as distance metric. We could apply more sophisticated metrics and/or machine learning algorithms such as SVMs to improve fication results. As our main goal in this work is to compare different ure descriptors and not machine learning algorithms, we have chosen e kNN approach.

al Feature Descriptors

e recognition system we describe in this paper starts building the pLSA age representation by describing each image with a number of feature f one kind. Then a vocabulary for that feature is computed as de- the previous section and a bag-of words representation is derived for ge. Local image features are often used in this context as they have the e of being more flexible than global image characterizations, while at time capturing more meaningful patterns than individual pixel values. predefined interest point at a specified scale (i.e., size of local neigh-, they describe the local image region surrounding the interest point y by a feature vector. There exists a large number of different types of ures, e.g. [11,12,13,14], each capturing a different property of a local gion and being more or less invariant to illumination, changes in view- d other image transformations. In the following we will use the term nd *descriptor* interchangeably.

vestigate the performance of the following three local feature descriptors ntext of the pLSA model:

1]**:** A SIFT feature for a detected interest point is computed by first g the orientation of the most dominant gradient. Then, relative to this on the gradient-based feature vector entries are computed from the lo- scale neighborhood. This is done by dividing the local neighborhood egions and subsequently accumulating the gradient magnitudes of each a local orientation histograms. The gradients are then weighted with an window centered at the interest point location. The entries of the ntation histograms form the entries of the, in our case, 128-dimensional ector. The vector is normalized to ensure invariance to illumination s. SIFT features are also invariant to small geometric distortions and ons due to location quantization. They are widely used in several com- on and pattern recognition tasks. Thus the results obtained with SIFT serve us as a baseline here.

ric blur [13]**:** The geometric blur feature vector computation is based ed edge channels, which in our work are computed by the boundary ctor proposed by Martin et al. [17]. A sub-descriptor is determined for e channel; the concatenation of all sub-descriptors forms the final geo- ur descriptor. In order to compute a sub-descriptor we collect the values e points in the neighborhood of the interest point. Sample points lie on c circles around the interest point. The outmost circle in this work has of 20 pixels. The distance between the 6 concentric circles decreases dratic manner. As twelve equally distributed sample values are taken circle the size of each sub-descriptor is 72 and thus the dimensionality tire feature vector is 288 when using four oriented edge channels. The

ilarity [14]**:** To derive the self-similarity feature for an interest point, called correlation surface is computed for the surrounding neighbor- compare a small image patch of size $x_1 \times x_1$ around the interest point larger surrounding image region of size $x_2 \times x_2$. In this work we choose nd $x_2 = 41$. Comparison is based on the sum of square differences be- gray values. The distance surface itself is then normalized and trans- nto a correlation surface, which in turn is transformed into a log-polar e system and partitioned into 80 bins (20 angles, 4 radial intervals). imum values in each bin constitute the local self-similarity descriptor. ing the descriptor vector ensures some invariance to color and illumi- anges. Invariance against small local affine and non-rigid deformations d by the log-polar representation; by choosing the maximal correlation each bin, the descriptor becomes insensitive to small translations.

vestigated feature descriptors are purely based on gray-scale images. ormance of scene classification, as considered in this work, is likely to by taking color into account (e.g. color SIFT [4]). As this may not be other content analysis tasks using probabilistic topic models such as cognition or image retrieval (because here categories might be defined rather than color), we do not consider color in this work.

mpute local features as described above at predefined interest points ssociated scale factor defining the size of the supporting image region he interest point. In order to be able to compare the different local rs, we will also analyze the behavior of the most common feature, the ture, for three different interest point detectors. We will pick the best ng detector for feature evaluation. The considered detectors are:

rence of Gaussian (DoG) detector [11]: Here a DoG pyramid is com- l. Interest points are defined as scale space extrema in the DoG pyramid are associated with its respective scale. Thus the DoG detector facili- scale invariant computation of the subsequent local feature descriptor supporting region size takes the scale factor into account. Note that in approach the number of interest points per image varies as it depends e structure and texture in each image.

e grid over several scales: We compute interest points on a dense grid spacing d between grid points in x- and y-directions and over several . As all images in our experiments are of the same size, the same number erest points is computed for each image.

sampling [13]: In this approach we require interest points to be located sitions of high edge energy. First we compute oriented edge channels ing a boundary detector [17]. Then all edge channels are thresholded ng only locations of high edge energy. Interest points are computed by omly sampling those locations. For random sampling all edge channels nsidered, nevertheless every position is selected at most once. Note that

e 1. Categories and number of images per category in the OT dataset

	1	2	3	4	5	6	7	8
e	coast	forest	highway	inside city	mountain	open country	street	tall building
s	360	328	260	308	374	410	292	356

Fig. 2. Sample images for each category in the OT dataset

perimental Evaluation

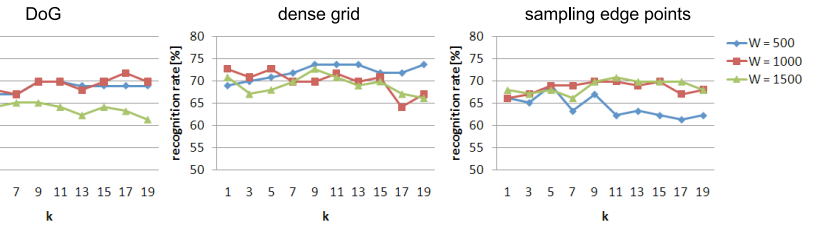
ental Setup: We use the OT dataset [18] to evaluate the three dif-erest region detectors and descriptors in the context of a scene classi-ask. The database consists of a total of 2688 images from 8 different egories. The number of images as well as examples for each category n in Table 1 and Fig. 2, respectively. On this dataset we perform im-fication by assigning each test image automatically to one of the eight s.

ide the images randomly into 1344 training and 1344 test images. We ubdivide the 1344 training images into a training and a validation set, 238 and 106 respectively. We used the validation set to find the best r configuration for the pLSA model. In the model we fix the number to 25 and optimize only the number of distinct visual words for the detectors/descriptors. A number of 25 topics has been shown to give a formance for this dataset [4].

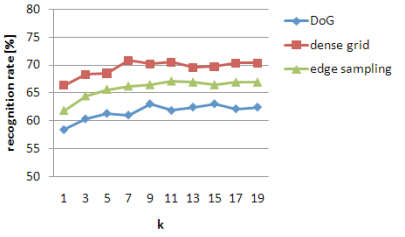
g determined the optimal number of visual words for the current detec-iptor combination we re-train the pLSA model with the entire training erging training and validation set. Final results are then computed on et and detector/descriptor performances are compared.

experiments, we will first analyze the suitability of three feature de-the scene classification task while holding the feature descriptor fixed. pick the best performing detector to evaluate the local descriptors.

Point Detectors: We select the frequently used SIFT descriptor for arison of the three detectors. Their parameters are set as follows: the between grid points is 5 pixels, resulting in about 5250 features per



ecognition rates on the validation set for the three different detectors over k of kNN for different numbers of visual words

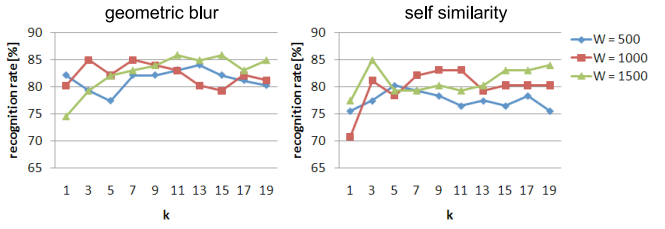


ecognition rates on the test set for three different detectors over k for kNN

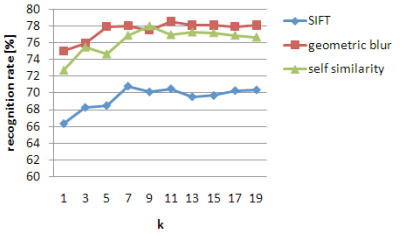
3 displays the resulting recognition rates on the validation set for numbers of visual words W for all three detectors over the parameter kNN algorithm. We observe that for the DoG detector, the dense grid over several scales, and the edge sampling detector $W = 1000$, $W = 500$ 1000 gives the best recognition results, respectively.

these parameter settings we train a pLSA model on the entire training ach detector type and fit the test set images to this model in order te a topic vector representation for all images. The comparison of the on results on the test set can be seen in Fig. 4. The dense grid detector ms the other detectors followed by random edge sampling.

ay be due to several reasons: Firstly, both the dense grid detector and om edge sampling algorithm compute more features per image than detector and also, they compute an equal number of features for each his may enable a better fitting of the pLSA model to the scene recogni- lem. Secondly, the interest points and regions computed by the dense r the entire image and thus the bag-of-words image representation also e entire image and not only regions close to edge pixels or scale-space A further reason might be that in a scene recognition task the repeata- exact positions and scales, as provided by the DoG detector, may be portant as in other tasks such as object recognition where one would atch only the exact subpart. The DoG detector offers in contrast to detectors scale invariance. Nevertheless this is also not as important



ecognition rates on the validation set for the two descriptors for different of visual words and k in kNN



ecognition rates on the test set over k for kNN for different local feature types

approximately the same scale. Note that the results are consistent with results [4], where a dense representation performed best, too.

Descriptors: The dense grid detector showed the best recognition nce in the evaluation above, thus we use this interest point detector bsequent comparison of local feature descriptors. First we determine opriate number of visual words in the pLSA model for each descriptor. already been done for the SIFT feature (see Fig. 3). Figure 5 depicts nition rates for different k in the kNN and different numbers of visual r the geometric blur descriptor and the self-similarity descriptor. The lts for both features are obtained using 1500 visual words.

th descriptors we train a novel pLSA model on the entire training set oute a topic vector representation for all training and test images. Then are the results of all local features, including SIFT, in Fig. 6.

be seen that both, geometric blur and self-similarity features outper- commonly used SIFT feature by more than 5%. Moreover the geomet- feature has a slightly better recognition rate, about 1% better, than imilarity feature, and the best recognition is achieved for $k = 11$ with It should be noted in this context, that a performance difference of statistically significant given the small OT dataset. Nevertheless, the arity descriptor is of lower dimensionality compared to the geometric ures: 80 vs. 288 dimensions. This lower dimensionality makes computa-

T (W=500, k=7)

	3	4	5	6	7	8
	13,33	0,00	5,56	9,44	2,78	1,11
	0,61	0,00	3,66	3,05	0,00	0,00
	76,15	2,31	1,54	5,38	6,92	2,31
	1,95	81,82	0,00	0,65	10,39	4,55
	2,14	0,00	66,31	14,44	4,28	0,53
	8,78	0,00	12,20	50,73	3,41	0,98
	1,37	8,22	4,11	0,00	82,88	2,74
	6,18	12,92	1,69	1,69	15,17	60,67

geometric blur (W=1500, k=11)

	1	2	3	4	5	6	7	8
1	78,33	0,56	7,78	0,00	0,56	12,78	0,00	0,00
2	0,00	93,29	0,00	0,00	1,83	1,83	3,05	0,00
3	16,15	0,00	70,77	2,31	1,54	4,62	4,62	0,00
4	4,55	0,00	0,00	81,82	0,00	1,30	3,90	8,44
5	0,53	4,81	4,28	0,00	79,14	7,49	3,74	0,00
6	21,46	4,39	5,85	0,00	6,83	59,02	2,44	0,00
7	0,00	0,00	2,05	7,53	1,37	0,00	85,62	3,42
8	0,00	1,69	1,12	8,43	0,56	0,00	3,93	84,27

self similarity (W=1500, k=11)

	1	2	3	4	5	6	7	8
1	66,11	1,11	9,44	0,56	2,78	17,78	2,22	0,00
2	0,00	90,24	0,00	0,00	2,44	3,66	3,66	0,00
3	9,23	0,00	73,08	0,77	6,92	3,85	6,15	0,00
4	1,30	0,00	0,65	82,47	0,00	0,00	5,84	9,74
5	4,28	6,42	3,74	0,53	70,05	6,95	8,02	0,00
6	9,27	5,37	3,90	0,49	4,88	74,63	1,46	0,00
7	0,00	2,05	1,37	3,42	1,37	0,68	91,10	0,00
8	0,00	0,56	0,00	10,11	0,00	0,00	8,99	80,34

onfusion tables for results on the test set for different descriptor types and a region detector. The numbers 1,2,...8 refer to the categories listed in Table 1.

ure. Thus, given the similar performance and the more than a magnitude mputational complexity over geometric blur, the self-similarity feature fered feature[1].

more detailed analysis of the results, the confusion tables for the best ng parameter settings for each descriptor are depicted in Fig. 7. In the tables it can be seen that there are some categories, such as *forest*, *y* and *street*, where all descriptors work almost equally well, showing a nce of over 80% and in the *forest* category achieving over 90% accuracy. noticed some confusions occur between closely related categories with sual appearance, e.g. *open country* and *coast*, *tall building* and *inside* ll as *mountain* and *open country*. In this cases, results might be further by including color.

rgest differences can be noticed in the category *tall building* where an about 20% smaller recognition rate than both other features. The c blur descriptor significantly outperforms SIFT and self similarity in ories *coast* and *mountain*, whereas the self-similarity feature performs he *open country* category.

we would like to examine the variance in performance due to random ion in both, the k-means clustering algorithm and the pLSA implen. Therefore we choose the parameter and feature setting of the best ng configuration so far (geometric blur descriptor, $W = 1500$, $k = 11$) at the scene classification experiment on the test set ten times, each puting the visual vocabulary and pLSA model with different ranalizations. The recognition rates range between 77.75% and 79.69% average value of 78.93% and a standard deviation of 0.58%. It can hat there are no large variations between different runs of the same nt.

mary it can be stated that for scene classification geometric blur outthe other features. In cases where fast computation is needed one should ider using the lower dimensional and faster-to-compute self-similarity hich only performs slightly worse than the geometric blur feature.

nclusion

ork we have studied the influence of the type of local feature descrip-
e context of pLSA based image models and a scene recognition task.
are three different local feature descriptors. Our results show that the
y used SIFT descriptor is outperformed by the two other feature de-
the geometric blur feature and the self-similarity features. Moreover
valuate three different local interest region detectors with respect to
ability in this task and we found that a dense grid detector over several
forms best. Future work could consist in adopting the best performing
rs to color.

nces

ann, T.: Unsupervised learning by probabilistic Latent Semantic Analysis.
Learn. 42(1-2), 177–196 (2001)
D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. J. Mach. Learn.
, 993–1022 (2003)
i, L., Perona, P.: A Bayesian hierarchical model for learning natural scene
ories. In: CVPR, pp. 524–531 (2005)
, A., Zisserman, A., Munoz, X.: Scene classification via pLSA. In: ECCV
)
as, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T., van Gool,
odeling scenes with local descriptors and latent aspects. In: ICCV, pp. 883–
005)
J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering
s and their location in images. In: ICCV (2005)
L., Fei-Fei, L.: Spatially coherent latent topic models for concurrent object
ntation and classification. In: ICCV (2007)
rd, K., Duygulu, P., de Freitas, N., Forsyth, D., Blei, D., Jordan, M.: Match-
ords and pictures. J. Mach. Learn. Res. 3, 1107–1135 (2003)
art, R., Slaney, M.: pLSA on large scale image databases. In: ICASSP (2007)
er, E., Lienhart, R., Slaney, M.: Image retrieval on large-scale image
ases. ACM CIVR, 17-24 (2007)
D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2),
0 (2004)
gie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using
context. PAMI 2(4), 509–522 (2002)
C.A., Berg, T.L., Malik, J.: Shape matching and object recognition using
stortion correspondences. In: CVPR (2005)
tman, E., Irani, M.: Matching local self-similarities across images and videos.
VPR (2007)
ajczyk, K., Schmid, C.: A performance evaluation of local descriptors.
27(10), 1615–1630
ster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete
via the EM algorithm. Journal of the Loyal Statistical Society B.39 (1977)
n, D., Fowlkes, C., Malik, J.: Learning to detect natural image boundaries
local brightness, color, and texture cues. PAMI 26(5), 530–549 (2004)