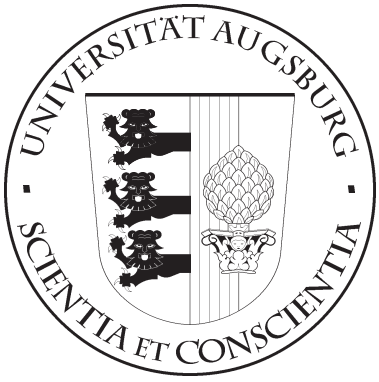
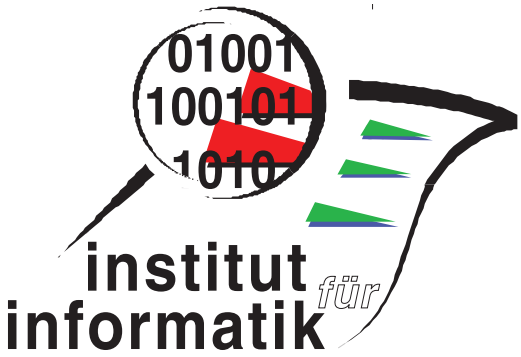
Universität Augsburg



Filtering adult image content with topic models

R. Lienhart, R. Hauke

Report 2009-10 Mai 2009



Institut für Informatik
D-86135 Augsburg

Copyright © R. Lienhart, R. Hauke
Institut für Informatik
Universität Augsburg
D–86135 Augsburg, Germany
http://www.Informatik.Uni-Augsburg.DE
— all rights reserved —

FILTERING ADULT IMAGE CONTENT WITH TOPIC MODELS

Rainer Lienhart

Lehrstuhl für Multimedia Computing
Universität Augsburg
Augsburg, Germany

Rudolf Hauke

Advanced US Technology Group, Inc.
1005 West Fourth Street
Carson City, Nevada, USA

ABSTRACT

Protecting children from exposure to adult content has become a serious problem in the real world. Current statistics show that, for instance, the average age of first Internet exposure to pornography is 11 years, that the largest consumer group of Internet pornography is the age group of 12-to-17-year-olds and that 90% of the 8-to-16-year-olds have viewed porn online. To protect our children, effective algorithms for detecting adult images are needed. In this research we evaluate the use of *probabilistic Latent Semantic Analysis* (pLSA) for this task. We will show that topic models based on pLSA can detect adult content with a correct positive rate of 92.7%, while only showing off a false positive rate of 1.9%. Even when using grayscale images only, a correct positive rate of 90.8% at a false positive rate of 2% can be achieved.

Index Terms— topic models, image classification, adult image content recognition, porn image detection

1. INTRODUCTION

Protecting our kids from being spammed with adult image content is – without doubt – a very pressing issue. In this paper we analyze how well recent concepts from image classification in general can be exploited for filtering adult content. Recently very successful approaches to image classification use topic models on visual words derived from salient descriptors of local image patches [12, 1, 10, 7]. The best-known topic model is the probabilistic Latent Semantic Analysis (pLSA) [8]. Thus, we will investigate how well image models based on pLSA can separate adult image content from normal image content. We will also investigate the usefulness of color information for this application domain and how much is lost if replying on grayscale images only.

Related Work: There are two major approaches to detecting porn images: Either (a) they focus on the text of the web pages accompanying the image to classify the image content or (b) they look inside the images and judge the content based on the amount of skin color pixels or skin texture pixels. Often some simple geometric constrains are applied, too. Examples of the first appraoch are [6, 9] and of the latter [5, 4]. Early approaches design their classification scheme based on manually tweaked heuristics, while latter approach use statistical classifiers such as Neural Networks or Support Vector Machines. However, none of them have used topic models yet with the single exception of [3]. Thus, there is a need to evaluate topics models for adult content recognition.

The paper is organized as follows. In Sec. 2 we introduce the overall probabilistic approach and describe the visual features in Sec. 3. It is followed by a discussion of the pros and cons of the approach in Sec. 4. Sec. 5 reports the experimental results, before we conclude the paper in Sec. 6.

2. APPROACH

We use a pLSA model to represent each image [12, 1, 10]. pLSA [8] was originally derived for text modeling, where words represent the elementary parts of documents. Building a pLSA model starts with representing the entire document corpus by a *term-document matrix* $[n(w_j, d_i)]_{M \times N}$. M indicates the number of documents in the corpus and i the associated document index variable, while N specifies the number of different words occurring across the corpus and j the associated word index variable. Each matrix entry stores the number of times a specific word w_j is observed in a given document d_i. Such a representation ignores the order of words in each document and is thus called a *bag-of-words* model.

In order to be able to apply this model to images, we need to define a visual equivalent to *words in documents*. Visual words are often derived from images by vector quantizing automatically extracted local region descriptors. In this work, a subset of local features extracted from training images are clustered by k-means clustering to derive the cluster centers as our visual vocabulary. Given the visual vocabulary we extract the local features from each image in the database and replace each with its most similar visual word. Similarity is defined as the closest word in the high-dimensional feature space. The word occurrences for each image are then counted, resulting in a term-frequency vector for each image document. These term-frequency vectors for each image constitute the term-document matrix. Note that any geometric relationship between the occurrences of different visual words in an image is disregarded since the term order is ignored.

Model: Given the term-document matrix, the pLSA uses

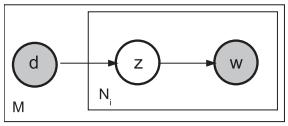


Fig. 1. Graphical representation of the pLSA model: M=# of images in database, N_i = # of visual words in image d_i, observable random variable w for the occurrence of a visual word and d for the respective image document, z = hidden topic variable.

a finite number of hidden topics to model the co-occurrence of visual words inside images across the image corpus. Each image is explained as a mixture of hidden topics. These hidden topics can be thought of as refering to objects or object parts. Thus we model an image as consisting of multiple object parts: For instance, an image of a beach scene consists of pieces of water, sand and people. Thus, assuming that every word w_j occurring in document d_i of the corpus is associated with a hidden (unobservable) topic variable z_k, the pLSA describes the probability of seeing word w_j in document d_i by the following model:

$$P(w_j, d_i) = P(d_i) \sum_k P(w_j|z_k)P(z_k|d_i) \qquad (1)$$

where $P(d_i)$ is the prior probability of picking document d_i, $P(z_k|d_i)$ the probability of selecting a hidden topic depending on the current document d_i (also referred to as the topic vector), and $P(w_j|z_k)$ the word distribution given a topic z_k. Fig. 1 shows a graphical representation of the pLSA model.

Learning and inference of topic models: We learn the probability distributions $P(w_j|z_k)$ of visual words given a hidden topic as well as the probability distributions $P(z_k|d_i)$ of hidden topics given a document completely unsupervised using the Expectation-Maximization (EM) algorithm [8, 2]. Probability distributions of new images that are not contained in the original training corpus are estimated by applying the EM algorithm to the unseen images to compute its topic distribution while keeping the learned word distributions conditioned on the topic $P(w_j|z_k)$ fixed. In our work, we compute the parameters of a pLSA model on the training data and then apply this model to unseen test data. We represent each image d by its associated topic vector $P(z|d)$ which gives us a very low-dimensional image representation.

Classification: For image classification the topic vectors of each unlabeled test image are classified by simple k-Nearest Neighbor (kNN) search through the labeled training images using the L1-norm as distance metric. Reasons for this choice are discussed in 4. The overall approach is visualized in Figure 2.

3. LOCAL FEATURE DESCRIPTORS

Based on our prior work [7] we compute two different local feature descriptors from each local region centered at (x, y):

SIFT [11]: The SIFT feature is computed by first calculating the orientation of the most dominant gradient. Then, relative to this orientation the gradient-based feature vector entries are computed from the local gray-scale neighborhood by dividing the local neighborhood into 4×4 subregions and subsequently accumulating the gradient magnitudes of each pixel into a local orientation histograms. The gradients are weighted with a Gaussian window centered at the interest point location. The entries of the 16 local orientation histograms form the entries of a 128-dimensional feature vector. The vector is normalized to ensure invariance to illumination conditions. SIFT features are also invariant to small geometric distortions and translations due to location quantization. They are widely used in several computer vision and pattern recognition tasks. Thus the results obtained with SIFT features serve us as a baseline.

Self similarity [13]: The self-similarity feature is computed by first calculating a so-called correlation surface for the surrounding neighborhood. We compare a small image patch of size $x_1 \times x_1$ around the interest point with a larger surrounding image region of size $x_2 \times x_2$. In this work we choose $x_1 = 5$ and $x_2 = 41$. Comparison is based on the squared L2-norm between the grayscale or color patches (C1R configuration). The distance surface itself is then normalized and transformed into a correlation surface, which in turn is transformed into a log-polar coordinate system and partitioned into 80 bins (20 angles, 4 radial intervals). The maximum values in each bin constitute the local self-similarity descriptor. Normalizing the descriptor vector ensures some invariance to color and illumination changes. Invariance against small local affine and non-rigid deformations is achieved by the log-polar representation; by choosing the maximal correlation value in each bin, the descriptor becomes insensitive to small translations.

Both features are originally defined for grayscale images only. An obvious extension to 3-channel color images is to derive at a given point $(x, y, scale)$ the base feature for each color channel independently in order to concatenate the three channel-specific feature vectors to a 384-dimensional SIFT vector or 240-dimensional self similarity vector. We label this configuration by C3R. For the self-similarity feature, we have the option to use the squared L2-norm on color images to computed the correlation surface, on which the 80-dimensional feature vector is computed (configuration C1R). We will see later in the experimental results that this is not only the mode that allows faster retrieval, but also the best mode.

Dense Sampling: We compute interest points on a dense grid with spacing d between grid points in the x- and y-directions and over several scales. As all images are scaled to the same length of the longest side, while preserving the original aspect ratio, the number of interest points computed for each image is about the same.

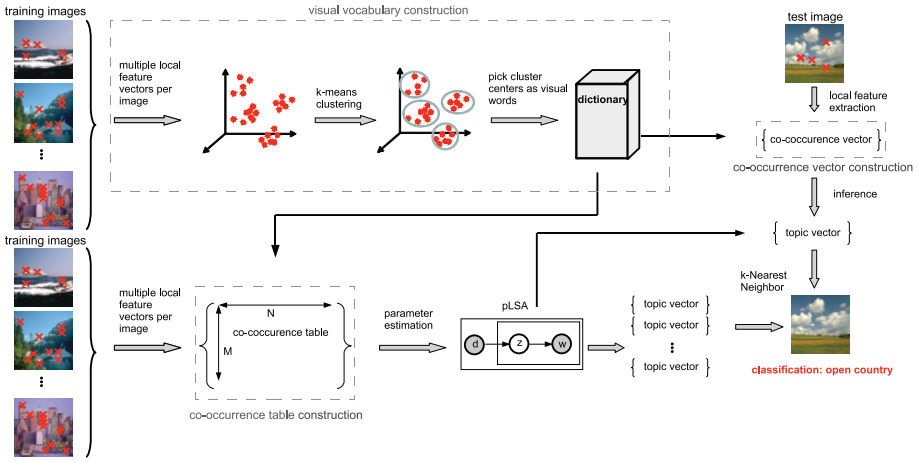


Fig. 2. Scene classification system based on a discrete pLSA model. Adult scenes are distinguished from everything else.

4. ATTRIBUTES OF THE PLSA APPROACH

This section explains reasons behind some of our design choices. Our image classification into porn vs. porn-free images is based on example-based k-NN classification and not on some discriminate learning algorithm such as Adaboot or SVM. Discriminate learning obviously would boost our already excellent performance numbers in Sec. 5. However, at the same time it would add inflexibility: Any time a new class of objectionable images needs to be filtered, a new training run would be required. In contrast, with the k-NN approach images that have been misclassified or images of a new objectionable image class can be added to the reference image set for the k-NN search at any time. Thus, in practice the adult filter can be updated easily to any kind of image content one wishes to filter. No retraining is required. Objectionable image content is so diverse that one can never assume to have a complete and representative sample set. Thus discriminate learning would not operate well in this domain due to frequently required retraining phases.

Given our example-based approach using pLSA has several advantages: Firstly, it compresses the high-dimensional document vectors (i.e., the word occurrence vector, 500 dimensions here) into a much smaller topic vector (50 dimensions here) by which each image is represented. This makes k-NN search much faster and scalable to large databases. Secondly, it is well-known that the smaller topic vectors produce better classification results compared to the larger document vectors (see [10]). Thirdly, pLSA not just enables adult vs. non-adult content filtering, but image search in general. Thus, general image search comes as a free lunch.

5. EXPERIMENTAL EVALUATIONS

Working with adult content is difficult in many ways: ethically and legally. Thus, we adopted a scheme that minimizes the time and the number of people who had to work with the access-restricted adult content. All parameter evaluations and optimizations were done with images of lightly dressed bikini models, which come as close as possible to adult content and sometimes even cross the border. However, these images do not require putting strict access restrictions into effect. Only in the very end we tested the overall system with the optimal parameters on a real adult content database to which only one authenticated person had access.

We tested the following color spaces: grayscale, hsv, hls, lab, luv, rgb, xyz, and xcrcb. For most of these color spaces the pixels can not only be represented by floating point numbers (denoted by "32f"), but also be range-compressed to a one unsigned byte representation (denoted by "8u").

Data Set 1 consists of 20,699 images from 18 different classes as listed in Table 2. Of these images 7,676 (containing 600 bikini images) were used for training and 13,023 (containing 512 bikini images) for testing. The accuracy for bikini vs. the other classes was normalized per class, so that the actual image count per category did not matter for the performance computation. The performance numbers for the various color spaces for the self-similarity feature and the SIFT feature are plotted in Fig. 3(a) and Fig. 3(b). As it can be clearly seen, the correct negative rates for both, the SIFT and the self-similarity feature, are very close to each other for the best performing color spaces. With the self-similarity feature correct negative rates of about 98.3% can be achieved in the color spaces *hls-8u-C1R*, *rgb-8u-C1R* and *rgb-8u-C1R*, while *lab-8u-C3R* reaches 98.1% with the SIFT features. Note, however, that the computation of the self-similarity feature is more than 3x faster to compute than SIFT, because the best performance was achieved with "C1R" and thus did not require computing 3 times the feature for each color channel independently ("C3R" mode). Instead the colors are considered early then calculating the correlation surface for the self-similarity feature, producing only a 80-dimensional vec-

Table 1. Comp. of data set 1 and 2 for the best color spaces.

color	data set	pos. rate	neg. rate	avg. rate
hls-32f-C1R	1	0.9133	0.9690	0.9412
hls-32f-C1R	2	0.8986	0.9776	0.9381
hls-8u-C3R	1	0.8901	0.9731	0.9316
hls-8u-C3R	**2**	**0.9266**	**0.9811**	**0.9538**
luv-8u-C3R	1	0.9114	0.9721	0.9417
luv-8u-C3R	2	0.9281	0.9626	0.9454
ycrcb-32f-C1R	1	0.9133	0.9690	0.9412
ycrcb-32f-C1R	2	0.8986	0.9776	0.9381
gray-8u-C1R	1	0.8540	0.9800	0.9170
gray-8u-C1R	2	0.9078	0.9798	0.9438

Table 2. Number of images per scene category.

scene cat.	#	scene cat.	#	scene cat.	#
airplanes	1074	forests	328	fields	410
beach	360	guitars	1030	streets	552
bikini	1112	homes	1000	storefronts	308
bottles	247	horses	170	skyscrapers	355
camels	346	motorbikes	826	faces	8499
cars	1281	mountains	374	people	2416

tor and thus adding only a very little run-time penalty.

With respect to the correct positive rate, the self-similarity feature clearly outperformed SIFT by achieving 91.1% for *hls-32f-C1R*, *luv-8u-C3R*, and *ycrcb-32f-C1R*. In contrast, SIFT in the best case of lab-32f-C3R can only achieve 88.9%. Thus, for our second data set with real adult images we will only consider the self-similarity feature on a reduced set of possible color space that have proven to be promising with the bikini model images.

Data Set 2 is identical to Data Set 1 except that the bikini images were replaced by 2,668 adult content images of which 600 were used for training and 2,068 for testing. Our goal is to see whether the results with the bikini images can be transferred to real adult content images. Table contrast the performance numbers for the three color spaces with the best average performance determined with the bikini images to the adult image set. As it can be clearly seen, the performance numbers are equvalent and sometimes even better for the adult images.

6. CONCLUSION

We have shown that current topic models are more than suitable for filtering images with adult content. A correct negative rate of 98.1% could be achieved at a correct positive rate of 92.7%. These performance numbers are way better than the results reported so far in the literature. Since topic models can easily be extended to incorporate information from other modes, even better performance can be expected in future.

7. REFERENCES

[1] A. Bosch, A. Zisserman, X. Munoz. Scene classification via pLSA. In ECCV, (2006).

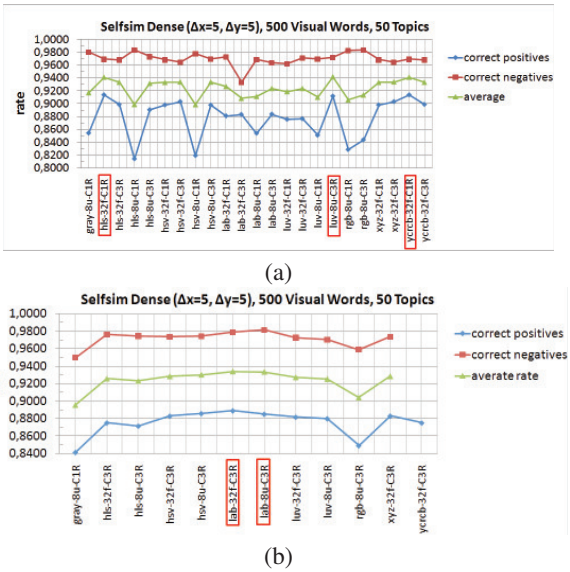


Fig. 3. The correct positive and correct negative rates are plotted against the various color spaces for (a) the self-similarity and (b) the SIFT features. The rates are averaged performance values for k={5,7,9,11,13,15,17,19} in the k-NN classification. The average rate curve is the average of the correct positive and negative rates.

[2] A. P. Dempster, N. M.Laird, D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Loyal Statistical Society, B.39, (1977).

[3] T. Deselaers, L. Pimenidis, H. Ney. Bag-of-Visual-Word Models for Adult Image Classification and Filtering In ICPR08, pp. 1–4, (2008).

[4] L. Duan, G. Cui, W. Gao, and H. Zhang. Adult Image Detection Method Base-on Skin Color Model and Support Vector Machine. ACCV2002, (2002).

[5] D.A. Forsyth, M. Fleek, and C. Bregler. Finding naked people. In Proc. of 4th European Conf. on Computer Vision, pp. 593–602, 1996.

[6] M. Hammami, Y. Chahir, L. Chen. WebGuard: Web based adult content detection and filtering system. IEEE/WIC Int. Conf. on Web Intelligence 2003, pp. 574-578, 2003.

[7] E. Hörster, R. Lienhart, M. Slaney. Image retrieval on large-scale image databases. ACM CIVR, 17-24, (2007).

[8] T. Hofmann. Unsupervised learning by probabilistic Latent Semantic Analysis. Mach. Learn., 42(1-2):177–196, (2001).

[9] Youngsoo Kim and Taekyong Nam. An efficient text filter for adult Web documents 8th Int. Conf. on Advanced Communication Technology (ICACT), 2006.

[10] R. Lienhart, M. Slaney. pLSA on large scale image databases. In ICASSP, (2007).

[11] D. G. Lowe. Distinctive image features from scale-invariant keypoints. IJCV, 60(2):91–110, (2004).

[12] P. Quelhas, F. Monay, J.-M. Odobez, D. Gatica-Perez, T. Tuytelaars, L.V. Gool.: Modeling scenes with local descriptors and latent aspects. In ICCV, pp. 883–890, (2005).

[13] E. Shechtman, M. Irani. Matching local self-similarities across images and videos. CVPR, (2007).