# Bag-of-Visual-Words Models for Adult Image Classification and Filtering

Thomas Deselaers[1]*, Lexi Pimenidis[2], Hermann Ney[1]*

[1] Human Language Technology and Pattern Recognition – [2] Security and Privacy Research
Computer Science Department, RWTH Aachen University, Aachen, Germany
E-mail: deselaers@cs.rwth-aachen.de

## Abstract

*We present a method to classify images into different categories of pornographic content to create a system for filtering pornographic images from network traffic. Although different systems for this application were presented in the past, most of these systems are based on simple skin colour features and have rather poor performance. Recent advances in the image recognition field in particular for the classification of objects have shown that bag-of-visual-words-approaches are a good method for many image classification problems. The system we present here, is based on this approach, uses a task-specific visual vocabulary and is trained and evaluated on an image database of 8500 images from different categories. It is shown that it clearly outperforms earlier systems on this dataset and further evaluation on two novel web-traffic collections shows the good performance of the proposed system.*

## 1 Introduction

Rating images according to their content is an important application area, with one main application in filtering network traffic to prohibit e.g. viewing pornographic material. One desired property of such a system is the possibility to dynamically change the content-type that is filtered to avoid the necessity of several such systems. Different clients might require differently strict content-filters (e.g. elementary schools or religious institutions might require different filters than universities or private employers). At home, people might want to enable such a filter over the day, when children are using the computer but disable it in the late evening [14]. Ideally, an pornographic image filter is created once and then the filter administrator can easily select which types of images he wants the filter to remove and which types of images are allowed.

In the literature, different porn image filtering techniques were presented: The detection of skin coloured areas is investigated in [10, 9], skin colour features are used in combination with other features such as texture features and colour histograms [7, 11, 15, 2, 9, 16].

Most of these systems build on neural networks or support vector machines as classifiers. In [14], some specialised features for porn image classification are presented and used in a retrieval/nearest neighbour classification scheme. The POESIA filter[1] contains an open source implementation of a skin-colour-based filter. Other approaches try to fuse textual and visual information from webpages in order to achieve better performance [8].

Recently the *bag-of-visual-words* (BOVW) models, which were initially proposed for texture classification [3, 13], have gained enormous popularity in object classification [4, 5] and natural scene analysis [6]. The BOVW models are inspired by the *bag-of-words* models in text classification where a document is represented by an unsorted set of the contained words. Analogously, here an image is represented by an unsorted set of discrete *visual words*, which are obtained by discretisation of local descriptors. The here presented method learns a task-specific visual vocabulary and employs a log-linear model to discriminate between different classes of content-type.

## 2 Porn Image Identification

For porn image identification, we follow the BOVW-approach, where images are represented as a histogram of visual words. The visual words denote local features extracted from the images and the vocabulary is learnt task-specifically from a training database.

### 2.1 Bag-of-Visual-Words Method

As local features, we extract image patches around difference-of-Gaussian interest points [12] which are scaled to a common size and then PCA transformed leaving 30 coefficients to reduce their dimensionality. The advantage of patches over e.g. SIFT features [12] is the straight-forward inclusion of colour information which clearly is important for the addressed task.

To create a visual vocabulary, we use the training algorithm for unsupervised training of Gaussian mixture models. This algorithm creates a set of $2^{\#splits}$ densities by iteratively splitting each existing density in the

[1] http://www.poesia-filter.org/

direction of its variance starting from a single Gaussian. The so learned visual vocabulary is able to capture frequently occurring patterns in the provided training data.

Given this model, for each local feature, the identifier of the closest density $\hat{c}$ is determined and a histogram over these identifiers is created:

$$h_c(X) = \sum_l \delta(c, \arg \max_{\hat{c}} \{\mathcal{N}(x_l|\mu_c, \Sigma_c)\}), \quad (1)$$

where $x_l$ are the local features extracted from image $X$, $h$ is the resulting histogram and $c$ enumerates the histogram bins and the corresponding densities. Thus, $h_c(X)$ gives the number of patches $x_l$ for which the $c$-th density is the one with the highest emission probability.

In the second phase, we use a discriminative classifier to determine the class of the images represented by the histograms. Here, we evaluate the use of support vector machines and log-linear models. For the support vector machine, we use the one-against-the-rest multiclass scheme, which allows to unify the decision rule for both approaches as follows:

$$r(h(X)) = \arg \max_{k'} \{g(h(X), k')\} \quad (2)$$

where the discriminant function $g(h(X), k)$ for the support vector machine is defined as

$$g_{\text{svm}}(h(X), k) = \alpha_k + \sum_{v \in \mathcal{S}_k} \alpha_v \mathcal{K}(h(X), v), \quad (3)$$

with $\mathcal{S}$ is the set of support vectors, $\alpha_c$ is the bias term for class $k$ and $\mathcal{K}$ is the kernel function. As the feature vectors in our experiments here are always histograms, we use the histogram intersection kernel[1]. The discriminant function of the LLM is defined as

$$g_{\text{llm}}(h(X), k) = \frac{\exp(\alpha_k + \lambda_k^t h(X))}{\sum_{k'} \exp(\alpha_{k'} + \lambda_{k'}^t h(X))}, \quad (4)$$

where $\lambda_k$ and $\alpha_k$ are the class specific parameters that are obtained using gradient descent and $\lambda_k^t h(X)$ denotes the scalar product of the histogram and the trained $\lambda_k$s.

## 2.2 Filtering Rules

To create a filter system from the classifier described above, first we define a set of categories, closely following [14], where the images are grouped into five different categories:

**class 0:** inoffensive images,

**class 1:** lightly dressed persons, might be offensive in very strict environments,

**class 2:** partly nude persons, might be objectionable in school environments,

**class 3:** nude persons, likely objectionable in many environments, and

**class 4:** porn images, probably offensive in most environments.

An example image for each of these classes (with blackened areas to keep the paper unobjectionable) is given in Figure 1 (classes 0 to 4 from left to right).
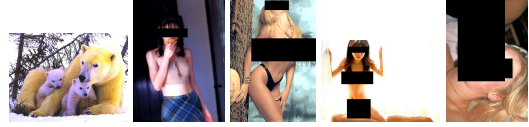


**Figure 1. Example images for the 5 different classes**

To give maximal flexibility to the administrator of the filtering system, the decision rule is defined as

$$r(h(X)) = \text{sgn}\left(\sum_{c=f}^{4} g(h(X), c) - \theta\right) \quad (5)$$

where an image is filtered, when $r(h(X))$ is $+1$ and allowed when it is $-1$. The parameter $f$ specifies the least objectionable class that should be filtered and $\theta$ is a threshold parameter that can be used to tune the false reject/false accept ratio according to the users needs. The chosen value for $\theta$ normally has to be chosen to find the best compromise between removing unobjectionable images (which happens frequently with small $\theta$) and not removing objectionable images (which may happen with large $\theta$). sgn is the signum-step function centred at 0.

The tuning of $\theta$ corresponds to changing the misclassification costs in Bayes' decision rule to minimise the expected cost of misclassification, and thus effectively allowing the filter administrator to decide about the amount of false positives (fp), i.e. images that are removed by the filtering software but should not, and the amount of false negatives (fn), i.e. images that should be removed by the filtering software but are not. This approach, which is based around a classifier discriminating five classes will in the following be denoted as '*joint*' classifier.

Another option to create a filtering system, is to create classifiers that are trained with known $f$-parameter. This method requires separate systems for different $f$-parameters. The resulting two-class classifiers can the directly be used to filter images without the need to use Eq (5) as the classes here directly correspond to 'objectionable' or not. Analogous, a thresholding parameter $\theta$ to tune the ratio of false positives and false negatives can be incorporated.

## 3 Experimental Results

To be able to compare our results with other published results, we use the database presented in [14]. The database consists of 8,500 images in total (approx. 1,700 per class), and we perform the experiments in five-fold cross-validation to keep training and test data separate. In informal experiments we evaluated the size of the visual vocabulary and found that the performance is clearly improved up to 2048 bins (11 splits) but is not much improved with more bins.

As described above, the evaluation of image filter-

ing/porn image identification, cannot be simply done by using the error rate. Instead, the number of unfiltered, but offensive images and the number of filtered, but inoffensive images is important (false negatives (fn) and false positives (fp)). Additionally, the different types of false negatives could be evaluated; e.g. in a setup where classes 2 to 4 are not-allowed, an image from class 2 which is accidentally allowed is probably not as offensive as an image from class 4, which passed the filter.

Results from the experimental evaluation of the methods are presented in Table 1 for different parameters $f$ with $\theta$ set to minimise the classification error on the training data. Table 1 (a) gives the false positive and false negative rates in percent for the SVM and the LLM. Both models perform similar, with slightly better classification of the SVM. The confusion matrices for these experiments are given in Table 1(b,c). In both confusion matrices, it can be observed that hardly any confusions between classes 0 and 4 happen, and that neighbouring classes, in particular classes 1 to 3, are confused more frequently.

In Table 2(a), comparison results to those in Table 1(a) are given, but instead of the joint classification approach defined by Eq (5), here binary classifiers were used. Again, the SVM slightly outperforms the LLM, but overall it can be observed that the joint approach is better than the binary classification task. This might be due to the additional knowledge encoded in the class hierarchy and the smaller intra-class variability in these experiments.

A comparison of our results to results presented in the literature is given in Table 2(b). The top block lists results on different setups using the AIRS system [14]. Note that these results are not strictly comparable to the results reported for our method, since we used only a subset of the data.

To demonstrate the power of the BOVW-approach over the skin colour models, we performed experiments, where we replace the histograms over visual words by skin colour histograms [10]. Results from these experiments are given in Table 2(c). It can be observed that the BOVW model clearly outperforms the skin colour features. Nonetheless, if we use skin colour features in combination with our visual vocabulary, a small improvement can be obtained as can be seen in Table 3(b).

An overview over most results presented here is given in Table 3(a): ROC curves of different experiments are given. The solid lines denote ROC curves for the experiments with LLM, the red dots denote the results given in [14] and the classification results for our experiments using the skin colour model from [10], respectively.

The results show that the automatic filtering of pornographic images can be done using a BOVW approach with better precision than using skin colour fea-

tures, but that a fusion of these approaches outperforms both of its components. Additionally, the overall precision of the filter depends on the filtering rules. It is much easier to create very strict (only class 0 allowed) or very slack filters (only class 4 prohibited) than it is to create filter with a more complex objective. The decisions between classes 1, 2, and 3 are much harder than the decision whether an image is class 0 or class 4.

An additional improvement is possible by combining skin colour models [10] with the BOVW models. To combine these two models, the skin colour histograms and the BOVW histograms are concatenated and the classifiers are trained on these augmented histograms. Results for these experiments are given in Table 3(b) and a comparison to the results in Tables 1(a) shows that the results are slightly better than the BOVW words model alone and much better than the skin colour model xalone. This shows that the BOVW model is able to capture nearly sufficient colour information that an additional skin colour model is not necessary.

**Experiments on Web Traffic Data.** Additionally to the experiments on the standard data set, we use two internal datasets of 1,000 images each. ==Dataset A was obtained by saving images that were routed through a proxy in a public anonymiser network, and Dataset B was obtained from the central proxy server of an internet provider.== We manually classified these 2,000 images according to the classification scheme described above. The traffic from the anonymiser network contains a much higher percentage of pornographic images than the normal web traffic, in particular the amount of images in the higher classes is increased. Results from the experiments on these datasets are given in Table 3(c). It can be observed that these datasets are much more difficult to filter than the standard dataset. Again, the $\theta$-parameter was chosen such that the classification error rate was minised to give a better impression of the systems performance. To create an actual filtering system, $\theta$ would be tuned to have a sufficiently low false positive rate.

## 4 Discussion and Conclusion

We have presented a porn-detection and filtering method based on the popular BOVW image classification model. ==The method allows for creation of a flexible content-filter that can be easily adapted for the users needs and clearly outperforms state of the art methods in this task on a standard task.== Integrating standard skin colour features into our system only led to a minor improvment, which demonstrates the general capability of the BOVW model to capture sufficient image information for this task.

**Table 1. (a) True positive (tp) and false positive (fn) rates for the SVM and the LLM experiments with a joint classifier. (b) Confusion matrix for the SVM experiment, (c) Confusion matrix for the LLM experiment.**
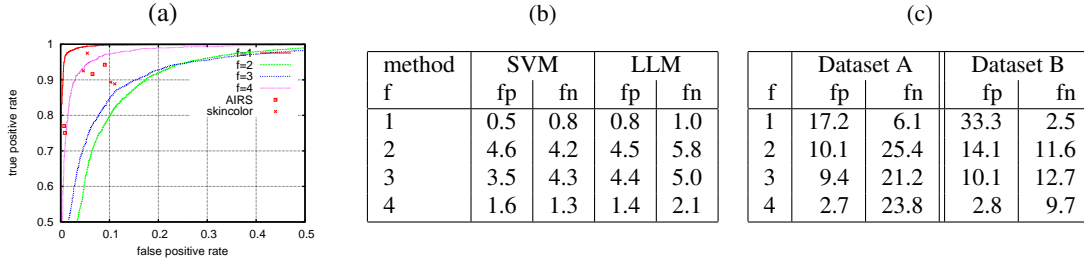
(a)

| $f$ | SVM fp | SVM fn | LLM fp | LLM fn |
|---|---|---|---|---|
| 1 | 0.5 | 0.8 | 1.2 | 0.8 |
| 2 | 4.4 | 4.6 | 6.5 | 4.3 |
| 3 | 3.4 | 4.7 | 5.1 | 4.6 |
| 4 | 1.6 | 1.5 | 2.0 | 1.7 |

(b)

| correct class | classified as 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 19.4 | 0.3 | 0.1 | 0.1 | 0.0 |
| 1 | 0.4 | 15.4 | 1.9 | 1.4 | 1.0 |
| 2 | 0.1 | 1.3 | 17.6 | 0.6 | 0.3 |
| 3 | 0.3 | 1.7 | 1.4 | 16.2 | 0.3 |
| 4 | 0.0 | 1.2 | 0.0 | 0.3 | 18.6 |

(c)

| correct class | classified as 0 | 1 | 2 | 3 | 4 |
|---|---|---|---|---|---|
| 0 | 18.8 | 0.4 | 0.2 | 0.4 | 0.1 |
| 1 | 0.4 | 14.1 | 2.0 | 2.1 | 1.4 |
| 2 | 0.2 | 1.4 | 16.9 | 1.1 | 0.4 |
| 3 | 0.2 | 2.0 | 1.3 | 16.1 | 0.4 |
| 4 | 0.0 | 0.9 | 0.1 | 0.4 | 18.5 |

**Table 2. (a) Results from the binary classification experiments for the porn identification task. (b) Comparison result from the literature, (c) Experiments with skin colour features and a LLM in joint/binary classification experiments.**

(a)

| $f$ | SVM fp | SVM fn | LLM fp | LLM fn |
|---|---|---|---|---|
| 1 | 0.5 | 0.6 | 0.7 | 0.8 |
| 2 | 3.5 | 7.6 | 5.1 | 7.7 |
| 3 | 5.6 | 4.3 | 5.6 | 6.5 |
| 4 | 2.1 | 1.3 | 2.1 | 2.2 |

(b)

| method | fp | fn |
|---|---|---|
| AIRS [14]* | 24.94 | 0.90 |
| AIRS [14]* | 8.39 | 6.53 |
| AIRS [14]* | 23.00 | 0.75 |
| AIRS [14]* | 5.75 | 9.04 |
| skin colour [10] in [14] * | 14.2 | 7.50 |

(c)

| method f | joint fp | joint fn | binary fp | binary fn |
|---|---|---|---|---|
| 1 | 5.5 | 2.5 | 3.6 | 3.2 |
| 2 | 11.1 | 11.1 | 9.2 | 11.7 |
| 3 | 10.3 | 10.7 | 12.6 | 9.2 |
| 4 | 4.7 | 7.5 | 8.5 | 5.0 |

* the experiments from [14] are not strictly comparable to our experiments as we use only a subset of the data they used and we use cross-validation.

**Table 3. (a) ROC curves for a selection of different porn filtering experiments. (b) Results for the experiments with BOVW model and skin colour features. (c) Results on the web traffic datasets.**

(a)



(b)

| method f | SVM fp | SVM fn | LLM fp | LLM fn |
|---|---|---|---|---|
| 1 | 0.5 | 0.8 | 0.8 | 1.0 |
| 2 | 4.6 | 4.2 | 4.5 | 5.8 |
| 3 | 3.5 | 4.3 | 4.4 | 5.0 |
| 4 | 1.6 | 1.3 | 1.4 | 2.1 |

(c)

| f | Dataset A fp | Dataset A fn | Dataset B fp | Dataset B fn |
|---|---|---|---|---|
| 1 | 17.2 | 6.1 | 33.3 | 2.5 |
| 2 | 10.1 | 25.4 | 14.1 | 11.6 |
| 3 | 9.4 | 21.2 | 10.1 | 12.7 |
| 4 | 2.7 | 23.8 | 2.8 | 9.7 |

# References

[1] A. Barla, F. Odone, and A. Verri. Histogram intersection kernel for image classification. In *ICIP*, vol. 2, pp. 513–516, Barcelona, Spain, Sept. 2003.

[2] A. Bosson, G. C. Cawley, Y. Chan, and R. Harvey. Non-retrieval: Blocking pornographic images. In *CIVR*, pages 50–60, London, UK, July 2002.

[3] O. G. Cula and K. J. Dana. Compact representation of bidirectional texture functions. In *CVPR*, pp. 1041–1047, Hawaii, USA, Dec. 2001. IEEE.

[4] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *ECCV*, Prague, Czech Republic, 2004.

[5] G. Dorko, C. Schmid, I. GRAVIR-CNRS, and F. Montbonnot. Selection of scale-invariant parts for object class recognition. *ICCV*, pp. 634–639, 2003.

[6] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, vol. 2, pp. 524–531, San Diego, CA, USA, June 2005. IEEE.

[7] D. A. Forsyth and M. Fleck. Finding naked people. In *ECCV*, vol. 2, pages 593–602, Cambridge, UK, 1996.

[8] W. Hu, O. Wu, Z. Chen, Z. Fu, and S. Maybank. Recognition of pornographic web pp. by classifying texts and images. *PAMI*, 29(6):1019, 2008.

[9] F. Jiao, W. Gao, L. Duan, and G. Cui. Detecting adult image using multiple features. In *Int. Conf. Info-tech and Info-net*, vol. 3, pp. 378–383, Beijing, China, 2001.

[10] M. J. Jones and J. M. Rehg. Statistical color model with applications to skin detection. *IJCV*, 46(1):81–06, 2002.

[11] K. Liang, S. Scott, and M. Waqas. Detecting pornographic images. In *ACCV*, pp. 497–502, Jeju Island, Korea, Jan. 2004.

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

[13] M. Varma and A. Zisserman. Classifying images of materials: Achieving viewpoint and illumination independence. In *ECCV*, vol. 3, pages 255–271, Copenhagen, Denmark, June 2002.

[14] S.-J. Yoo. Intelligent multimedia information retrieval for identifying and rating adult images. In *Int. Conf. KES 2004*, vol. 1 of *lnai 3213*, pp. 164–170, Wellington, NZ, 2004.

[15] W. Zeng, W. Gao, T. Zhang, and Y. Liu. Image guarder: An intelligent detector for adult images. In *ACCV*, pp. 198–203, Jeju Island, Korea, Jan. 2004.

[16] H. Zheng, M. Daoudi, and B. Jedynak. Blocking adult images based on statistical skin detection. *ElCVIA*, 4(2):1–14, 2004.