AUTOMATIC DETECTION OF CHILD PORNOGRAPHY USING COLOR VISUAL WORDS

Adrian Ulges, Armin Stahl

German Research Center for Artificial Intelligence (DFKI),
D-67663 Kaiserslautern, Germany
e-mail: {adrian.ulges, armin.stahl}@dfki.de

ABSTRACT

This paper addresses the computer-aided detection of child sexual abuse (CSA) images, a challenge of growing importance in multimedia forensics and security. In contrast to previous solutions based on hashsums, file names, or the retrieval of visually similar images, we introduce a system which employs visual recognition techniques to automatically identify suspect material. Our approach is based on color-enhanced visual word features and a statistical classification using SVMs. The detector is adapted to CSA material in a training step.

In collaboration with police partners, we have conducted a quantitative evaluation on several datasets (including real-world CSA material). Our results indicate that recognizing child pornography is a challenging problem (more difficult than the detection of regular porn). Yet, while skin detection – a popular approach in pornography detection – fails, our approach can achieve a prioritization of content (equal error $11 - 24\%$) to improve the efficiency of forensic investigations of child sexual abuse. Examples illustrate that the system employs color cues as key features for discriminating CSA content.

Index Terms— child pornography detection; content-based image retrieval; visual recognition

1. INTRODUCTION

Over the last years, the spread of child pornographic data has increased at alarming rates. Though multiple efforts are being taken to combat this spread – ranging from the prevention of cyber grooming[1][2] over hotlines for reporting child abuse[3][4] to cooperations with internet service providers (1) – more and more child sexual abuse (CSA) material is distributed. For example, in Germany alone, the number of investigations in the area has grown from $1,500$ (2000) to $8,832$ (2007).

In parallel, the multimedia explosion, accompanied by high-bandwidth internet and cheaper storage devices, has led to a rapid growth of personal image and video collections. In this context, the efficient identification of child pornography within (potentially huge) multimedia databases has become a vital issue, not only to child protectors (targeted at banning child porn from the internet), but particularly to prosecutors confronted with the forensics analysis of suspect image and video material. Here, investigators face the difficult challenge of detecting child-pornographic material in large-scale databases (often $100,000$s, sometimes even millions of images) and under considerable time pressure. Correspondingly, it has been reported that prosecutions are abandoned because confiscated material cannot be analyzed in time.

Current solutions to detect CSA images employ file hashes like MD5 sums to match seized material with databases of known child pornography maintained by the police. This approach allows for a rapid, automatic detection of known child pornographic content, but requires the target images to be bit-identical to known CSA material – not only does this fail in case of small modifications such as re-encoding and resizing, but it is of no help for detecting novel material (which appears on the web constantly).

This paper follows a different approach: based on a statistical classification of a pictures' texture and color, our system estimates a *score* indicating whether the given image is child-pornographic. Our approach employs visual words (2) with color-enhanced DCT descriptors for representing images. The resulting feature vectors are fed to a Support Vector Machine (SVM) classifier (3) discriminating CSA images from non-CSA material. Though this approach cannot be expected to reach the accuracy of a careful manual investigation, it is valuable in a semi-automatic setting, where we can prioritize images or filter the "most suspicious" content for manual investigation. This has already been applied extensively for the detection of general pornographic content (4; 5; 6) – our work, however, is (to the best of our knowledge) the first to investigate visual recognition for *child* pornography detection.

We present a quantitative evaluation of our system on a variety of datasets (including real-world CSA material as well

[1] SAFE network – http://www.safenz.org/

[2] INSAFE network – http://www.saferinternet.org/web/guest/home

[3] http://www.missingkids.org

[4] https://www.inhope.org/

978-1-61284-350-6/11/$26.00 ©2011 IEEE

as Flickr photo stock, standard research datasets, and web pornography). Experiments were conducted in collaboration with police partners in the project FIVES[5], and include comparisons with a baseline using skin detection, a common approach for the identification of regular pornography (7; 8). Our results indicate that CSA detection is a challenging problem and more difficult than identifying regular pornography. Yet, with equal error rates in the range of $11 - 24\%$, CSA detection has the potential to prioritize content and significantly reduce the manual effort of forensic investigations in the area.

2. RELATED WORK

Beyond a matching on the basis of hash sums – which is fast and simple but does not detect modified or new CSA material – a variety of other methods have been developed.

Network-centric Child Porn Detection: Some efforts have been targeted at the identification of child pornography in computer networks. Shupo et al. (9) detect CSA material in network traffic by a statistical classification on packet level. The applicability of this approach to previously unseen material, however, has not been validated. Other approaches, like the project MAPAP ("Measurement and Analysis of P2P Activity Against Paedophile Content") focus on peer-to-peer file sharing networks. Here, CSA material is identified based on suspicious file names (10) or by modeling user activity (11). In contrast to this, we employ an images' *content* as a complementary information source that is indicative even in case of weak file naming (e.g., `0001.jpg`).

Content-based Image Retrieval: Approaches from *content-based image retrieval* (CBIR) (12) can be used to detect CSA material using a *query-by-example* strategy: given a questioned sample image, visually *similar* (but not necessarily identical) CSA material is retrieved. This approach has been integrated in several tools used in the forensic area (13; 5; 14), and has proven very useful for identifying images that are themselves unknown but come from well-known shoots, series, or locations (15). The approach has also been extended to the video domain, as in the commercial tool Videntifier Forensic[6] or in the EU project *i-dash* (16).

Visual Recognition: Visual recognition techniques are concerned with automatically recognizing objects (17), object categories (18), or even general semantic concepts (19) in images and videos. Over the last years, strong improvements have been achieved by so-called *patch-based* methods, which describe images as collections of local regions of interest. Particularly, novel approaches have been developed for the robust detection and description of interest regions, which substantially improves the robustness of recognition (e.g., (20; 17; 21; 22)). Our system builds on these modern patch-based approaches.

Pornography Detection: One visual recognition problem with particular relevance to our work is the detection of regular pornography. Here, a frequently used approach is to perform a statistical analysis on the basis of skin color detection. Forsyth et al. (23) matched the detected skin regions with human bodies by applying geometric grouping rules. Wang et al. (24) applied nearest neighbor classification to skin areas, achieving a speed-up by a fast filtering of icons and graphs. Jones and Rehg focused on an accurate detection of human skin by constructing RGB color histograms from a large-scale dataset of segmented training images (7). Porn detection is also of interest for web content filtering: for example, Rowley et al. used Jones' skin color histograms in a system installed in Google's Safesearch (8). Finally, closest to this work is a patch-based porn detection approach by Deselaers et al. (4), which extracts local features by Difference-of-Gaussian interest point detection, describes them with their PCA transformation, and quantizes them with a codebook of patch categories (or *visual words*). Our system follows a similar approach and applies it for the detection of CSA material.

3. APPROACH

We formulate the identification of CSA images as a two-class classification problem: for each questioned picture, we estimate a boolean random variable C indicating the presence ($C = 1$) or absence ($C = 0$) of child pornography. To do so, the content of the image is represented by a numerical feature vector $x \in \mathbb{R}^d$, which is fed to a statistical classifier (trained previously on a number of labeled training images). This classifier estimates a probability (or *score*) $P(C = 1|x)$.

3.1. Color-enhanced Visual Words

We adopt the frequently used *bag-of-visual-words* feature representation (2) for our system, an approach that has recently been shown to give strong results in a variety of visual recognition tasks such as object recognition (25) or concept detection (26). The approach has also been applied successfully for the detection of general pornographic content (4), which renders it a promising candidate for CSA detection.

The key idea of the approach is to describe images as collections of local interest regions. These are discretized using vector quantization, obtaining clusters of visually coherent *patches* referred to as *visual words* (in reminiscence to the well-known "bag-of-words" representation from textual information retrieval (27)). An illustration is given in Figure 1, which displays patches from the same visual word in the same line. Obviously, patches within a visual word are visually correlated, and can often be associated with coherent objects, like "faces" or "plants". We expect that the occurrence of certain visual words provides powerful hints for the presence of offensive (or even child-pornographic) material.

[5] http://fives.kau.se

[6] http://www.identifier.com

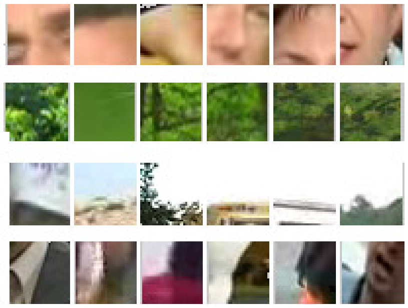

Fig. 1. Color-enhanced visual words: the patches in each line (corresponding to the same visual word) share a similar appearance and tend to show similar object parts. Histograms of these visual words are fed to an SVM classifier.

Our implementation extracts local interest regions by a regular sampling of overlapping rectangular patches of size 14×14 pixels at steps of 5 pixels (images are scaled to a width of 250 pixels). The resulting patch areas are described by applying the Discrete Cosine Transform (DCT) in YUV color space, and selecting 78 low-frequency coefficients (36 for luminance and 21 for each chroma channel). This setup was validated before to give a superior performance to other local features (such as SURF (20)) in a porn detection scenario (28). The resulting patches are vector quantized using a visual codebook of prototypes learned by a K-means clustering over a large-scale set of training images (we use $K = 2,000$, which can be considered a common choice (29)). Each patch is matched to its closest cluster center (or *visual word*), and the number of occurrences of each visual word in the image is counted. The resulting histogram is used as feature vector for classification with an Support Vector Machine (SVM). A χ^2 kernel was used, whereas the parameters C' (cost of training sample misclassification) and γ (kernel smoothness) were optimized using grid search.

3.2. Baseline: Skin Detection

To compare our system with a standard baseline, we employ skin detection, a frequently used approach in pornography recognition. We evaluate a setup similar to Jones' method (7), which models the distributions of RGB colors c in skin regions, $P(c|\mathbf{s})$, and background, $P(c|\neg\mathbf{s})$. For this, RGB color histograms are used, which were previously estimated from a set of segmented training images. For each pixel (with color c), a "skin probability" is estimated using Bayes' rule:

$$P(\mathbf{s}|c) = \frac{P(\mathbf{s}) \cdot P(c|\mathbf{s})}{P(\mathbf{s}) \cdot P(c|\mathbf{s}) + P(\neg\mathbf{s}) \cdot P(c|\neg\mathbf{s})},$$

with the prior tuned manually to $P(\mathbf{s}) = 0.2$. Repeating this for each pixel results in a *skin probability map* (see Figure 2). After binarization with a threshold of 0.5, size normalization, and some morphological post-processing, we obtain skin regions, from which four simple features are extracted: the average skin probability, the ratio of skin pixels (before and after morphological processing), and the size of the largest skin region. These features are fed to an SVM for classification (3) (using an RBF kernel). Just like for the visual words approach, parameters were fitted using a grid search.

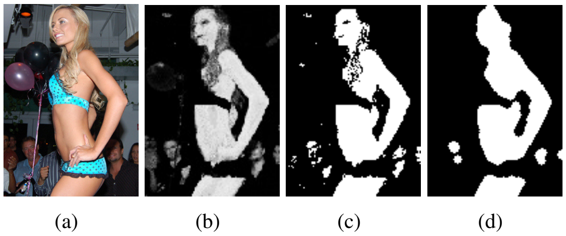

Fig. 2. As a baseline, we employ skin detection. Given an image (a), a skin probability map is estimated (b), binarized (c), and post-processed (d). From this information, simple statistics are extracted and fed to an SVM classifier.

4. EXPERIMENTS

We evaluate the recognition performance of our system by measuring the accuracy of detecting real-world CSA material within several other kinds of image content.

1. **CSA** – Our target material consists of $20,000$ CSA pictures collected by police partners in the project FIVES. The features described in Section 3 were extracted on police sites (we did not have access to the image material itself).

2. **Porn** – This dataset contains $4,248$ images showing regular pornography, acquired by a random crawl of pornographic websites and a manual filtering (removing logos, etc.). We use this material to compare CSA detection with the detection of regular pornography, and for testing the detection of child porn *within* regular porn.

3. **Flickr** – This dataset consists of $5,000$ inoffensive consumer photographs downloaded from the web portal Flickr. The pictures were acquired over several months by iteratively requesting the most recently uploaded images, giving an unbiased sample of Flickr content.

4. **Corel** – This is a subsample of $4,198$ inoffensive images randomly drawn from the Corel dataset (30), a well-known standard benchmark in image retrieval research.

5. **Web** – A dataset of $2,752$ inoffensive images crawled from popular non-pornographic websites, giving a representative mix of logos, graphics, portraits, and other photographs.

Table 1. Results of regular pornography ("Porn vs. ...") and child pornography ("CSA vs. ...") detection. Our color visual words approach outperforms skin detection, particularly when it comes to detecting child pornography.

	equal error rate (%)	
	skin detection	color viswords
Porn vs. Flickr	9.5 ± 0.7	$\mathbf{9.2 \pm 0.3}$
Porn vs. Corel	8.3 ± 0.2	$\mathbf{6.0 \pm 0.5}$
Porn vs. Web	14.1 ± 0.4	$\mathbf{9.7 \pm 1.0}$
CSA vs. Flickr	28.5 ± 0.5	$\mathbf{21.8 \pm 0.3}$
CSA vs. Corel	26.3 ± 1.2	$\mathbf{11.2 \pm 0.7}$
CSA vs. Web	32.3 ± 0.7	$\mathbf{13.8 \pm 0.4}$
CSA vs. Porn	27.1 ± 0.6	$\mathbf{24.0 \pm 1.2}$
CSA vs. All	38.1 ± 0.7	$\mathbf{21.5 \pm 0.2}$

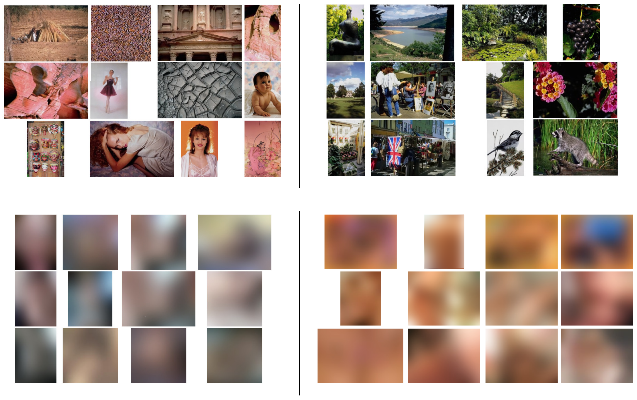

Fig. 3. Non-CSA samples from the experiments "CSA vs. Corel" (top) and "CSA vs. Porn" (bottom, pornographic samples blurred). Content on the left shows the highest CSA scores, content on the right the lowest.

We obtain a variety of test cases by mixing target material with content from the other datasets (e.g., "CSA vs. Flickr"). In each case, $2,000$ images ($1,000$ per class) were sampled for both training and testing, and detection accuracy was averaged over 5 runs of random resampling.

CSA detection vs. Pornography Detection: Table 1 shows equal error rates for different combinations of target and background material. We see that error rates for detecting regular pornography are significantly lower than for child porn: our system ("color viswords") as well as the baseline ("skin detection") achieve a reliable detection of regular pornography (with error rates in the range of $6 - 14\%$), whereas the visual words approach is more accurate. This confirms earlier results on pornography detection (4). The situation is different when it comes to detecting child porn. Here, error rates are significantly higher (in the range of $11 - 24\%$), indicating that identifying CSA material is a more challenging problem than detecting regular pornography. This is also illustrated by the ROC curves in Figure 4.

Comparison with Baseline: A second important observation is that – when comparing skin detection and visual words for CSA detection – our approach gives significant performance improvements, ranging from 6.7% (Flickr) to 18.5% (Web). This indicates that – while skin detection is a suitable approach for regular pornography (7; 8) – it should not be the first choice when it comes to CSA detection.

Generalization to Different Series: Our CSA and porn datasets may contain *series* (i.e., clusters of images taken at the same location and displaying similar objects), and obviously detection becomes easier when using images from the same series for training. Therefore, we also tested the system when forced to generalize to completely new material (this was achieved by manually sorting images such that training and test material were guaranteed to come from disjoint series). Results are illustrated in Figure 4 (dashed lines, "cross-series"). We see that performance drops occur in the range of 0.3% (Web) to 4.9% (Corel). Yet, though the generalization to unknown series poses an even more difficult challenge, our system still works in general.

Detection in Different Genres: Finally, we observe that detection accuracy varies with the genre of the content in which we detect CSA images. For example, CSA detection in Corel images works well (equal error 11%): as Figure 3 indicates, our system can reliably discard outdoor and landscape scenes in the Corel set (whereas color appears to be a vital source of information). On the other hand, the most difficult challenge is to discriminate CSA material from regular pornography (Figure 4, bottom right). Here, detection is far from reliable, with equal errors in the range of 24% and higher. Finally, when mixing all Non-CSA genres in equal shares ("CSA vs. All"), our system achieves a intermediate error of 21.5% (whereas skin detection fails short of handling the diversity of material [EER=38.1%]). Overall, our results indicate that CSA detection is a difficult challenge (though a ranking of content can be achieved), and that our color-enhanced visual words approach performs a much more accurate detection of CSA material than skin detection.

Examples: Examples are illustrated in Figure 3, where we display test images with highest and lowest CSA score. Obviously, the system makes strong use of color information, particularly when distinguish CSA material vs. regular porn (where the system has learned paler skin and relatively dark backgrounds to be discriminative features of CSA content).

5. CONCLUSION

We have presented a visual recognition system for the detection of child pornographic image material, based on color-enhanced visual word features and SVM classification. Our results indicate that the automatic detection of child sexual abuse (CSA) material is a much more difficult challenge than

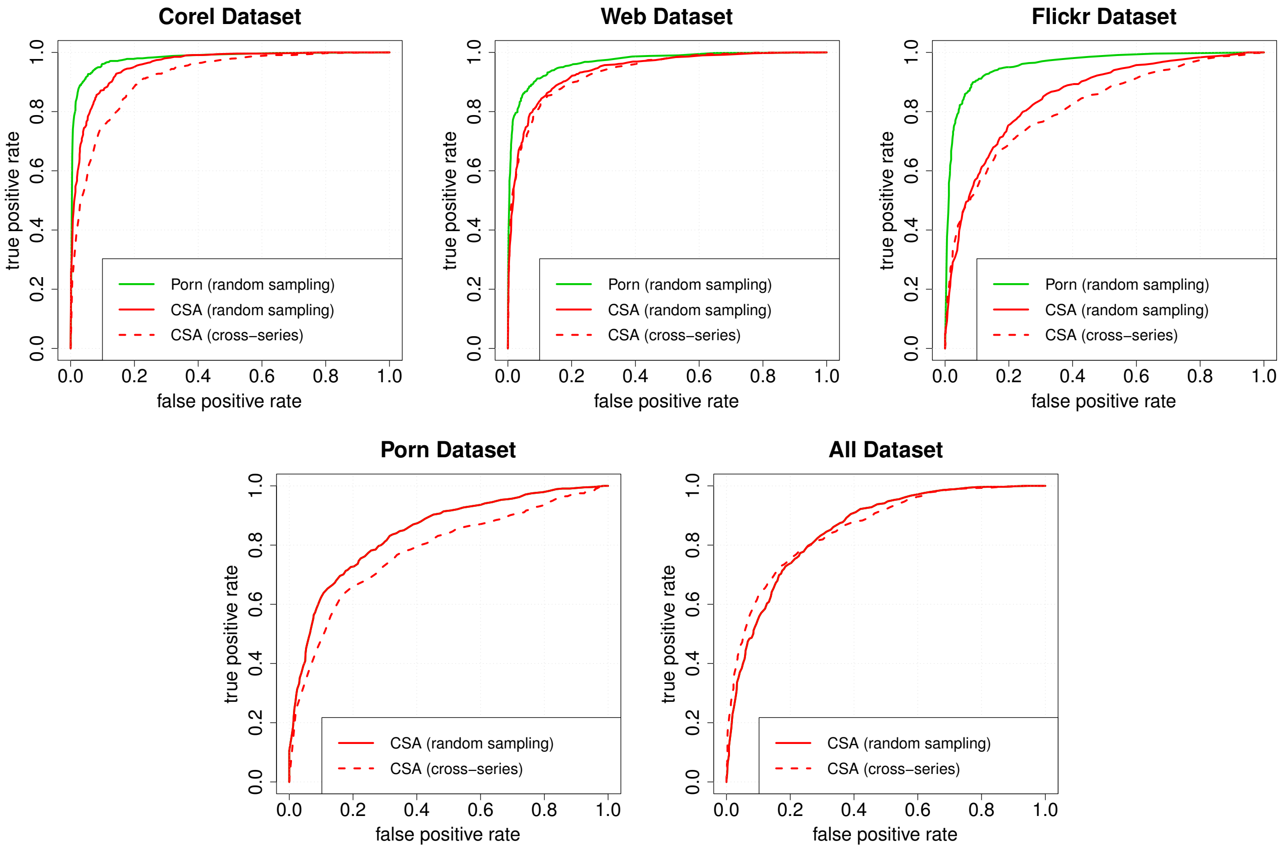

Fig. 4. As these ROC curves indicate, the detection of regular pornography (green) is more accurate, CSA detection (red) more difficult. Also, when comparing random sampling (solid) with the "cross-series" experiments (dashed), generalizing to series of previously unknown material is more challenging. When comparing different datasets, we see that CSA detection works well on the "Web" and "Corel" datasets, but is less accurate on "Flickr". The most difficult challenge is to detect CSA material within pornographic content (equal error 24%).

identifying regular pornography. However, while a mere detection of skin regions (as frequently applied for pornography detection) is not sufficient, our approach achieves a higher accuracy such that at least a prioritization of content becomes possible.

Our system automatically learns pale skin and dark, grayish background to be visual key features for identifying child pornography. It does not locate persons or identify them as children (though we observed a certain tendency of the system to detect facial close-ups). This raises the question whether our approach can be combined with a face detection and age estimation (31), which – though a challenging problem by itself – poses an interesting alternative. As both approaches focus on independent aspects of child pornographic material, their combination might lead to strong improvements.

From a practical perspective, we are currently integrating the evaluated technologies into a framework developed in the FIVES project. Here, CSA detection will become part of an integrated tool for the digital investigator, which we hope to form the basis for a prioritized, interactive, and efficient search for CSA material[7].

References

[1] EuroISPA, "Effectively Fighting the Online Distribution of Child Sexual Abuse Material," available from http://www.euroispa.org/ (retrieved: Oct 2010), September 2010.

[2] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *Proc. Int. Conf. Computer Vision*, Oct. 2003, pp. 1470–1477.

[3] C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition," *Data Min. Knowl. Discov.*, vol. 2, no. 2, pp. 121–167, 1998.

[7] This work was supported by the EU Safer Internet Programme, project FIVES (SIP-2008-TP-131801)

[4] T. Deselaers, L. Pimenidis, and H. Ney, “Bag-of-Visual-Words Models for Adult Image Classification and Filtering,” in *Proc. Int. Conf. Pattern Recognition*, December 2008, pp. 1–4.

[5] “LTU Engine,” available from http://www.ltutech.com/en/products/ltu-engine-2 (retrieved: June 2010).

[6] “ZiuZ Visual Intelligence,” available from http://www.ziuz.com/ (retrieved: June 2010).

[7] M. J. Jones and J. M. Rehg, “Statistical Color Models with Application to Skin Detection,” *Int. J. Comput. Vision*, vol. 46, no. 1, pp. 81–96, 2002.

[8] H. Rowley, Y. Jing, and S. Baluja, “Large Scale Image-Based Adult-Content Filtering,” in *Int. Conf. Comp. Vis. Theory and Applications*, February 2006, pp. 290–296.

[9] A. Shupo, M. Martin, L. Rueda, A. Bulkan, Y. Chen, and P. Hung, “Toward Efficient Detection of Child Pornography in the Network Infrastructure,” *Int. J. on Computer Science and Information Systems*, vol. 1, no. 2, pp. 15–31, 2006.

[10] M. Latapy, C. Magnien, and G. Valadon, “First Report on Database Specification and Access including Content Rating and Fake Detection System,” Tech. Rep., Measurement and Analysis of P2P Activity Against Paedophile Content (MAPAP) Project, 2008.

[11] J.-L. Guillaume, M. Latapy, C. Magnien, and G. Valadon, “Content Rating and Fake Detection System,” Tech. Rep., Measurement and Analysis of P2P Activity Against Paedophile Content (MAPAP) Project, 2009.

[12] A. Smeulders, M. Worring, S. Santini, and A. Gupta R. Jain, “Content-Based Image Retrieval at the End of the Early Years,” *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

[13] “imgSeek internet presence,” available from http://www.imgseek.net/ (retrieved: June 2010).

[14] “Netclean Analyze,” available from http://www.netclean.com/ (retrieved: June 2010).

[15] M. Ali and J. Hofmann, “An Extensive Approach to Content Based Image Retrieval Using Low- & High-Level Descriptors,” Tech. Rep., Master’s Thesis, IT University of Gothenburg, Sweden, 2006.

[16] “I-Dash Project Homepage,” available from http://www.i-dash.eu/ (retrieved: June 2010).

[17] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints,” *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[18] J. Ponce, M. Hebert, C. Schmid, and A. Zisserman, *Toward Category-Level Object Recognition*, Springer-Verlag New York, Inc., 2007.

[19] C. Snoek and M. Worring, “Concept-based Video Retrieval,” *Foundations and Trends in Information Retrieval*, vol. 4, no. 2, pp. 215–322, 2009.

[20] H. Bay, T. Tuytelaars, and L. van Gool, “SURF: Speeded Up Robust Features,” in *Proc. Europ. Conf. Computer Vision*, May 2006, pp. 404–417.

[21] K. Mikolajczyk, “A Performance Evaluation of Local Descriptors,” in *Proc. Int. Conf. Computer Vision and Pattern Recognition*, June 2003, pp. 257–263.

[22] P. Roth, “Survey of Appearance-based Methods for Object Recognition,” Tech. Rep. ICG-TR-01/08, Computer Graphics & Vision, TU Graz, 2008.

[23] M. M. Fleck, D. A. Forsyth, and C. Bregler, “Finding Naked People,” in *Proc. Europ. Conf. Computer Vision*, 1996, vol. 2, pp. 593–602.

[24] J. Z. Wang, G. Wiederhold, and O. Firschein, “System for Screening Objectionable Images using Daubechies’ Wavelets and Color Histograms,” in *IDMS ’97: Proceedings of the 4th International Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, 1997, pp. 20–30.

[25] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, “The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results,” October 2008.

[26] K. van de Sande, T. Gevers, and C. Snoek, “A Comparison of Color Features for Visual Concept Classification,” in *Proc. Int. Conf. Image and Video Retrieval*, July 2008, pp. 141–150.

[27] D. Lewis, “Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval,” in *Proc. Europ. Conf. Machine Learning*, Apr. 1998, pp. 4–15.

[28] C. Jansohn, “Statistical Classification of Image Content for Visual Information Filtering,” diploma thesis, University of Kaiserslautern (available from http://madm.dfki.de/teaching), 2009.

[29] E. Nowak, F. Jurie, and B. Triggs, “Sampling Strategies for Bag-of-Features Image Classification,” in *Proc. Europ. Conf. Computer Vision*, May 2006, pp. 490–503.

[30] H. Müller, S. Marchand-Maillet, and T. Pun, “The Truth about Corel - Evaluation in Image Retrieval,” in *Proc. Int. Conf. on Image and Video Retrieval*, July 2002, pp. 38–49.

[31] C. Küblbeck, T. Ruf, and A. Ernst, “A Modular Framework to Detect and Analyze Faces for Audience Measurement Systems,” in *GI Jahrestagung*, 2009, pp. 3941–3953.