Pornography Detection in Video Benefits (a lot) from a Multi-modal Approach

Adrian Ulges, Christian Schulze, Damian Borth, and Armin Stahl
German Research Center for Artificial Intelligence (DFKI) GmbH
Kaiserslautern, Germany
{adrian.ulges, christian.schulze, damian.borth, armin.stahl}@dfki.de

ABSTRACT

We address the challenge of detecting pornographic content in video streams. On offensive material crawled from different pornographic websites and non-offensive clips from YouTube (a total of 500 hours of video), we first study a compressed-domain activity descriptor based on MPEG motion compensation vectors. We show that the approach offers an interesting alternative but generalizes poorly between videos compressed with different codecs, a problem that can be overcome to some extent by adding noise to the image data prior to video compression.

Our main contribution is an evaluation that benchmarks the above motion-based descriptor as well as three other widely used features (audio-based MFCC features, skin color detection, and visual words). Here, we show that a multi-modal approach is a key strategy for an accurate detection or adult content: A combination of the different features gives considerable improvements in accuracy, reducing equal error by 36–56% compared to the best uni-modal system.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Indexing

General Terms

Algorithms, Measurement, Experimentation

Keywords

Pornography Detection, feature fusion, compressed-domain motion descriptors

1. INTRODUCTION

Due to the increasing spread of cameras and cell phones, higher internet bandwidth, and new media sharing services such as social networks, image and video collections have recently experienced a rapid growth that is commonly referred to as the *multimedia explosion*. On the one hand, this raises the challenge of granting users efficient access to material of interest (with applications such as search, browsing, and recommendation). On the other hand, however, some kinds of material may be unwanted, harmful, or even illegal, such as violence, political extremism, and offensive content.

The focus of this work is on the detection and filtering of pornographic content using an automated image and video analysis. One major application area for this technology is user protection on the web: Here, search engines like Google or Bing or on-line content sharing portals like YouTube and Flickr are already filtering adult material to avoid displaying it to the user. In this context, content-based image and video analysis serves as an important complement to other signals such as text-based classification or user flagging [5, 16]. Other applications of pornography detection include child protection and the detection of illegal content, such as child sexual abuse in forensic investigations [21].

Pornography detection is of particular interest for video data, which is often of enormous volume and is extremely time-consuming to analyze manually. This is particularly true as pornographic content may not cover the whole timeline of a video but may only appear at certain points in time (illegal video snippets might even be hidden inside longer video streams).

In this paper, we address the challenge of detecting pornographic content within video streams. Thereby, we take a multi-modal perspective: While previous work has focused on the analysis of one or two modalities in the input stream (such as the image content, motion, or audio), we present a quantitative study that compares *all* of these features – and their combination – on a large-scale diverse dataset of 500 hours of video crawled from the web. Second, we study a compressed-domain activity descriptor that has previously been suggested for pornography detection [6]. In particular, we evaluate the approach when generalizing between different input video formats. We experience strong overfitting here, an effect that can be reduced to some extent by adding image noise prior to transcoding the input video.

2. RELATED WORK

In the following, we discuss related work on offensive image and video detection that covers the different feature modalities targeted later in this paper. For image data, the most frequently used approach is the localization of skin color, followed by a description of the image by the quantity and shape of the detected skin regions. Jones and Rehg estimate the "skin probability" of a pixel by matching with

AMVA'12, November 2, 2012, Nara, Japan.

its color with skin and background color models learned on a large-scale image dataset [7]. Forsyth et al. [4] match the detected skin regions with human body parts by applying geometric grouping rules. Rowley employs skin detection in a large-scale system developed at Google [16] and combines it with other (more generic) features such as the image size, the graylevel entropy, and the number and size of detected faces. More recently, Deselaers et al. [1] investigated *bag-of-visual-words* features (which can be considered state-of-the-art in a variety of other recognition scenarios [3, 17]). The approach describes the image as a collection of vector-quantized local patches and was demonstrated to outperform skin color detection. In our evaluation, we will include both a skin detection approach and a variant of visual words features.

For video data, a common approach is to extract representative keyframes from the video stream and then apply image-based techniques: Lee et al. [10] used skin detection and color histograms, Kim et al. [8] a shape description of skin areas. Other methods are targeted at joining in audio analysis as an additional feature: Zuo et al. [22] employ Mel Frequency Cepstral Coefficients (MFCCs), a frequently used feature in speech processing [14]. In a similar fashion, Liu et al. [12] demonstrate improvements by combining visual features with *audio words*, i.e. vector-quantized features derived from the audio stream. Kim and Kim [9] use audio features based on the Radon transform, which is demonstrated to give moderate improvements over MFCCs. Rea et al. [15] detect periodic patterns in a video's audio signal as an indicator of sexual sounds such as moaning. We will include vector-quantized MFCC features (or *audiowords*) in our evaluation.

Finally, *motion* has been investigated as an additional modality. Again, periodicity detection has been employed, which can be based on an analysis of the motion signal's autocorrelation [18] or on a spectral analysis using *periodograms* [2]. As an alternative, Jansohn et al. [6] investigated *motion histograms* derived from MPEG motion compensation vectors, and demonstrated that the approach performs superior for detecting pornography in the wild (frequently, pornographic motion patterns were not found to be strictly periodic).

Overall, while previous work employs only a single feature modality or at most a combination of two (typically audio and static imagery [15, 22]), our main contribution is a combination of four of the most prominent features (namely, skin detection, visual words, MFCC features, and motion histograms), including a critical assessment of their accuracy in a real-world setting.

3. APPROACH

Our setup is illustrated in Figure 1: We perform a regular segmentation of the video stream into time windows of 4 seconds. For each of these time windows, we extract multi-modal features from the audiovisual content, including (1) motion histograms derived from the video's MPEG motion compensation vectors, (2) a description of the detected skin color and (3) visual words, both extracted from the center frame of the time window, and (4) vector-quantized MFCC features derived from the corresponding audio signal. Each of these features is fed to a statistical classifier (in previous experiments, Support Vector Machines [SVMs] were found superior to other options), and the resulting classification scores are combined to a joint multi-modal result in a *late fusion* step. In the following, we first describe the single features used (Sections 3.1 – 3.4), followed by an outline of statistical classification and classifier fusion (Section 3.5).

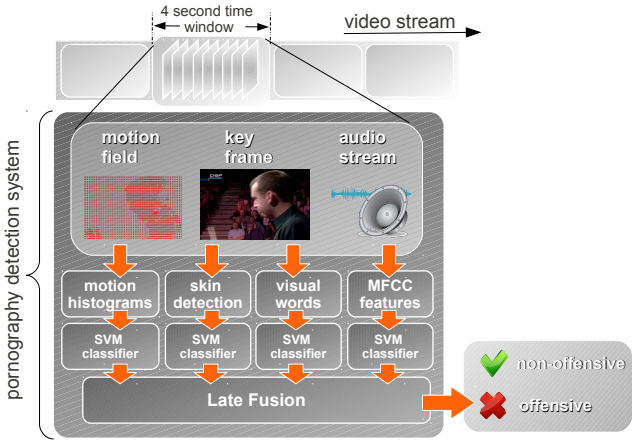


Figure 1: Our approach combines different feature description techniques for pornography detection in video, namely motion histograms, skin detection, visual words, and audio-based MFCC features. For each modality, features are extracted and a classifier is applied. Finally, uni-modal recognition results are combined in a late fusion.

3.1 Motion Histograms

We employ *motion histograms* that were first introduced for automatic video tagging [20] and have been applied for pornography detection by Jansohn et al. [6]. This feature is based on motion compensation vectors, which are estimated by video codecs for compression purposes during the transcoding process and are embedded in the video stream. These motion vectors can be extracted rapidly by a partial decompression of the video (our non-optimized implementation achieves a throughput of over 500 frames per second). The resulting motion fields indicate shifts of 16×16 image blocks between subsequent frames in the video, and can thus serve as an indicator of motion.

To compute a feature representation, video frames are divided into 4×3 subwindows, and for each subwindow all motion vectors (v_x, v_y) occurring over the 4-second time window in the respective region are combined in a 2D histogram. The histogram discretizes both components of the motion vector into 7×7 non-regular bins in the range $[-20, 20]$. By concatenating the 4×3 window-wise histograms, we obtain a 588-dimensional activity descriptor (which is normalized to sum 1).

Figure 2 gives an illustration of this feature, showing two video scenes and their respective motion histograms. For both scenes, we see 4×3 subwindows. In each one a motion histogram is displayed, with the center denoting the motion vector $(0, 0)$ (i.e., the absence of motion). The first scene (showing a video blog) displays little activity, such that the distribution of motion vectors is strongly peaked at the $(0, 0)$ motion vector (there is a white dot in the center but only few other motion vectors). The second scene shows pornographic content: We observe strong motion, partially due to camera shaking (note that intentionally no global motion compensation is applied), partially due to repetitive sexual motion patterns (as in the bottom left of the scene).

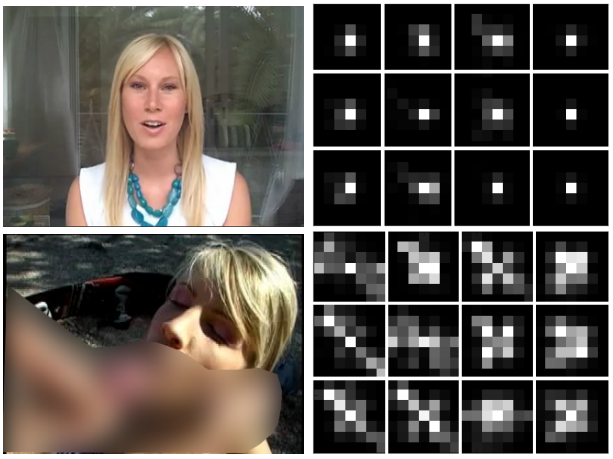


Figure 2: Two sample motion histograms: While the first video shows an almost static scene, the second one displays strong motion due to camera shaking and sexual activity.

3.2 Audio Features

We represent the audio stream using the well-known Mel Frequency Cepstral Coefficient (MFCC) features [13, 14]: After resampling the audio to a mono stream at 22,050 KHz, we extract MFCC features[1] and vector-quantize them to *audio words*. An MFCC feature is extracted every 8 milliseconds from a sliding time window of 16 milliseconds width (this parameter was optimized beforehand). After applying a Hamming window function, the audio snippet is fed to the Fourier transform. Over the resulting spectrum, a frequency histogram is computed (whereas the frequency range typically associated with human voice is emphasized by using a logarithmic Mel scale). This histogram is finally DCT-encoded, obtaining a 13-dimensional descriptor of each short-term audio snippet. We obtain 500 MFCCs for each 4-second time window, which we vector-quantize to 1,000 clusters pre-trained using K-Means. Similar *audiowords* features have successfully been used in pornography detection before [12].

3.3 Skin Color Detection

We also include two features based on static imagery, which we apply to the center frame of each 4-second time window. As a first option, we test skin color detection, a frequently used technique for detecting pornographic content. We follow the approach by Jones and Rehg [7]: Models of skin color ($P^{skin}(c)$) and non-skin color ($P^{non-skin}(c)$) are pre-trained on the COMPAQ image database of manually segmented pictures. These models are stored as 32^3-dimensional histograms in RGB color space. For each pixel in the frame (with color c), we estimate a probability

$$P(skin|c) = P^{skin}(c)/(P^{skin}(c) + P^{non-skin}(c)).$$

The resulting *skin probability map* (SPM) is post-processed using a morphological refinement, connected component analysis, and adaptive thresholding to obtain a skin segmentation mask (SSM). We store the mean intensities, as well as the center and variance of skin mass in both the SPM and SSM to obtain a 14-dimensional descriptor.

[1]using the implementation from the YAAFE library: http://yaafe.sourceforge.net/

3.4 Visual Words

While skin color detection can be a valuable and fast approach towards pornography detection, the resulting features are inherently limited as skin detection is sensitive to changes of illumination and the resulting features do not capture scene context nor variations in lighting and skin tone. Therefore, we apply a second image-based approach named *visual words* that gives a more holistic representation of the observed scene. This approach can be considered state-of-the-art in other recognition tasks such as object category recognition [3] or concept detection [17].

Our implementation is inspired by the one of Deselaers et al. [1]: After scaling the image to fit a square of 250^2 pixels, we extract patches of 8×8 pixels at regular steps of size 5. The resulting patches are color-transformed to YUV space, a Discrete Cosine Transform (DCT) is applied on each color channel, and we store 36 low-frequency components from the Y channel and 21 from each chroma channel. The resulting 78-dimensional patch descriptors are vector-quantized to a codebook of 2,000 clusters trained by K-Means.

3.5 Inference and Feature Combination

We combine the four above modalities in a weighted sum late fusion: For each feature a separate SVM classifier is trained (using RBF kernels for skin features and χ^2 kernels for visual words, audio features, and motion histograms). SVM meta-parameters (i.e., cost terms C and kernel smoothness γ) are estimated using a cross-validated grid search. Classification scores are mapped to probabilities using the sigmoid fitting proposed by Lin et al. [11]. The resulting probabilities are combined using a weighted sum:

$$P(porn|X) = \sum_{f \in \{\text{skin, viswords, audio, motion}\}} w_f \cdot P^f(porn|X), \quad (1)$$

whereas X denotes the 4-second time window to classify. Weights w_f are learned by a cross-validated grid search.

4. EXPERIMENTS

This section describes quantitative experiments benchmarking the accuracy of pornography detection when using the different multi-modal features described in Section 3, as well as their combination. We first describe the experimental setup in Section 4.1 (including the datasets used, preprocessing, and choices of various parameters). Section 4.2 will describe a first set of experiments with the *motion histogram* approach, particularly focusing on its invariance properties with respect to the codec of the input video. After this, Section 4.3 provides results of the individual features, and Section 4.4 discusses their combination into a joint decision.

4.1 Setup

We conduct our experiments on a dataset of web video clips crawled in winter 2011/2012 from *YouTube* (for non-offensive content) and from the two pornographic websites *redtube.com* and *pornhub.com*. Material from the pornographic websites was downloaded via "most recent" feeds (with the intention to sample a representative mix of content). From YouTube, material was crawled by text searches

Table 1: Equal error rates (%) of pornography detection when using each modality individually. We conduct two kinds of experiments: in-domain (training and testing on the same website) and cross-domain (generalizing between websites). It can be seen that motion histograms generalize poorly, which can be explained by their dependence on the input video codec.

	in-domain		cross-domain	
	pornhub → pornhub	redtube → redtube	pornhub → redtube	redtube → pornhub
motion histograms**	**10.70**	**12.20**	**21.80**	**33.90**
motion histograms (with noise)	–	–	17.70	26.50
skin detection	13.83	11.31	12.50	11.32
visual words	11.23	9.20	11.70	12.11
audio features	20.20	20.10	20.20	19.40

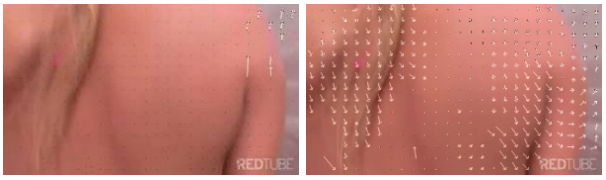


Figure 3: Motionfields estimated with different XViD settings, illustrating that motion compensation vectors heavily depend on the codec used and its setting.

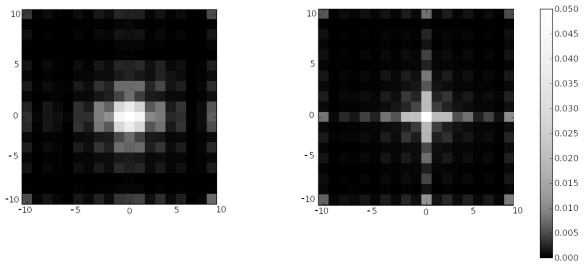


Figure 4: The distribution of motion vectors averaged over several hundred videos from redtube (center) and pornhub (right). Although all videos have been transcoded to the same format before recording motion, a significant difference in the motion statistics can be observed.

with 233 diverse search terms we used in previous work *[anonymous]*, with the resulting content including videoblogs, sports, interviews, comedy, etc. Overall, our dataset contains about 500 hours of video (1,000 pornographic clips with an average length of 19 minutes, and 2,300 YouTube clips with an average length of 5 minutes).

All videos (which come in different formats depending on the website) were transcoded using the XViD codec. Out of the resulting clips, 20-second snippets were sampled randomly, each consisting of five 4-second time windows (for the pornographic clips, we sampled from the second half of each clip, which we found more likely to show sexual activity). Training and classification were applied on the basis of these time windows (as illustrated in the beginning in Figure 1), and the classification score for each 20-second snippet was obtained using a simple averaging over its time windows.

Training sets of 4,000 video snippets were randomly sampled (2,000 positive ones and 2,000 negative ones) as well as test sets of 1,500 snippets (500 positive ones and 1,000 negative ones). It was made sure that no content from the same video was mixed between training and testing. This procedure was repeated 5 times, obtaining randomized training and test "folds". All results reported in the following are averaged over these 5 random iterations. The accuracy of detection is measured using equal error rate and ROC curves.

As we are interested in how well approaches generalize between different formats of input videos, we perform "in-domain" experiments (where training and testing happen on content from the same porn website) as well as "cross-domain" ones (where detectors are trained on one porn website but applied on the other). Background material was always sampled from YouTube.

To avoid training on the testing data, fusion weights for combining the different approaches (see Equation (1)) were learned using a 20-fold cross-validation.

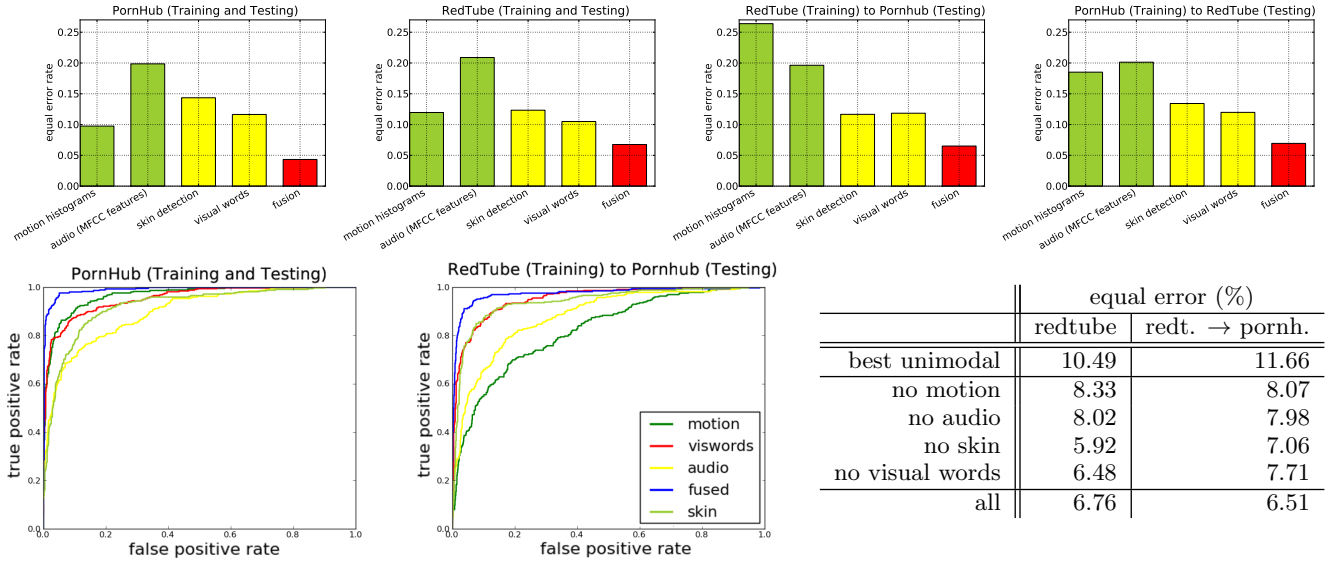
4.2 Experiment 1: Motion Histograms

Our first experiment focuses on the motion histogram feature described in Section 3.1. In particular, while previous work has indicated that motion compensation vectors can be a strong discriminator for pornographic content [6], our focus is on their generalization between different video codecs.

The reason for this is that video codecs estimate motion vectors for compression purposes, not to describe the actual scene motion. Different video codecs follow different motion estimation strategies (often a variation of block matching (e.g., [19]) and depending on the parameters of the codec motion may look quite differently. An example is illustrated in Figure 3: The same frame (transcoded with different settings of the XViD codec) shows very different motion fields – while in the left example motion estimation is omitted for most macroblocks (in which case motion is simply set to $(0, 0)$), on the right a more accurate motion field has been estimated.

To overcome this problem, we transcode all input videos with the same codec before extracting motion features, enforcing all content to undergo the same motion estimation[2]). Using this setup, we conducteded a first experiment evaluating motion-based classifiers. Results are illustrated in Table 1 (first row, marked with **). We see that – when testing on the same porn website as training on (*in-domain*), motion histograms give an acceptable performance: For pornhub,

[2]This transcoding was done using XViD, whereas the following parameters were found to ensure a sufficient image quality and motion estimation: `vhq=3`, `bitrate=700`, `me_quality=3`, `noqpel`, `nogmc`



	equal error (%)	
	redtube	redt. → pornh.
best unimodal	10.49	11.66
no motion	8.33	8.07
no audio	8.02	7.98
no skin	5.92	7.06
no visual words	6.48	7.71
all	6.76	6.51

Figure 5: Equal error rates (**top**) and ROC curves (**bottom left**) when fusing different modalities for pornography detection in video. In all cases, the combination of multiple modalities gives significant improvements in accuracy over the best uni-modal system. **Bottom right**: Equal error when using only one feature, when using all but one, and when using all features.

they are even the best feature (equal error 10.7%). However, when generalizing to a different website, error increases rapidly to 21.8% or even 33.9%. An explanation is given in Figure 4, which displays the distribution of motion vectors (averaged over several hundred videos) for content from redtube (left) and pornhub (right). We observe significant differences in the distribution of motion, even though all input videos have been transcoded to the same format. The explanation for this are *block artifacts*: Input videos often display a certain blockiness, with smooth intensities within macroblocks and edges at the block boundaries. When transcoding these videos, motion estimation is biased to align with the given block boundaries, i.e. the estimated motion is driven towards the motion in the original video.

To overcome this problem, we add *Gaussian noise* to the frames of the input videos prior to motion estimation. This reduces the effect of block boundaries and thus biases motion estimation less towards the original motion vectors, i.e. the distribution of motion vectors changes by transcoding. This is illustrated in Table 1 – when using the same motion vector histograms but adding Gaussian noise, error in the cross-domain experiments is reduced to some extent (from 21.8 to 17.7% for redtube and from 33.9 to 26.5% for pornhub).

4.3 Experiment 2: Evaluation of the Single Modalities

Quantitative results for the single modalities are illustrated in Table 1. Results for motion histograms have already been discussed in Section 4.2, with a poor generalization in the cross-domain settings that can be overcome to some extent by adding image noise. We see that the best overall performance is provided by the image-based features, with equal error remaining relatively stable when training on a different website than the one testing on. Thereby, the visual words approach gives a slightly better accuracy than skin detection in 3 of 4 cases. Audio features appear to be the weakest individual modality, with an equal error rate in the range of 20% – here, an in-depth inspection of results indicated that offensive videos did not necessarily come with pornographic audio signals (for example, simply because music was overlayed).

4.4 Experiment 3: Multi-modal Combination

Finally, we benchmark the combination of all four features, whereas for the cross-domain experiments the motion variant with image noise added prior to transcoding was used. Results are illustrated in Figure 5: We see that in all cases the combination of multiple modalities comes with significant improvements in accuracy compared to the best uni-modal system. These improvements range from 36% (redtube, 10.5% → 6.8%) to 56% (pornhub, 9.8% → 4.3%).

To assess how much the individual features contribute to a successful multi-modal system, we also measured combination results when leaving out each of the individual modalities (weighting the other modalities equally). Results are illustrated in Figure 5 (bottom right). Interestingly, it can be seen that leaving out motion histograms or audio decreases accuracy more than leaving out skin detection or visual words. This indicates that – even though audio and motion give a low accuracy when used individually – they contribute the most to a successful combined system by bringing in complementary aspects not covered by the other features.

Figure 6 provides an in-depth inspection of sample results. In each column, videos are displayed for which the multi-modal approach gives the *strongest improvements* over using *one particular* feature (i.e., for each feature we display videos for which the feature alone gave a poor result). False positives are displayed (top) as well as false negatives (bottom). We see that *skin detection* and *visual words* are attracted by scenes with large amounts of skin colored regions (like sand one a beach (b), or close-ups of hands (f)). Also, skin detection fails in case of unusual lighting (c, d) and visual words in case of significant blue areas in the scene (g,h), which the

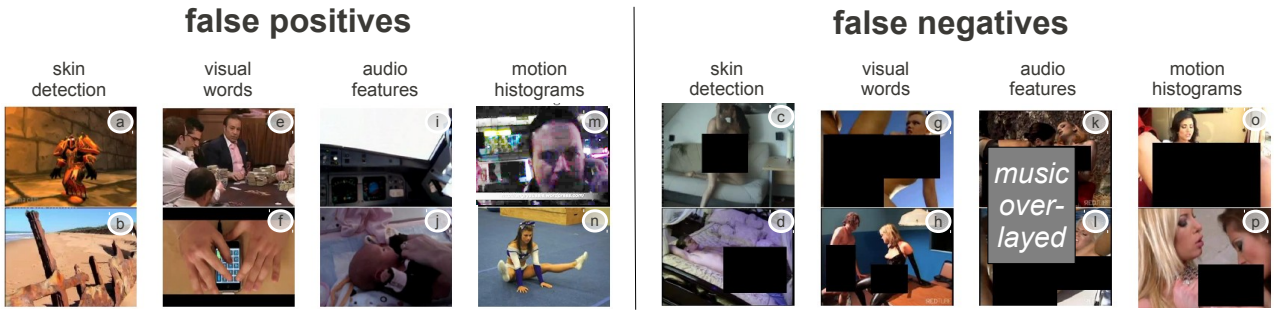


Figure 6: Sample detection results when using different feature modalities (in each column, we illustrate samples for which one of the individual features gave a poor result, either a false positive (top) or a false negative (bottom)).

approach takes as an indicator of outdoor landscape scenes, ranking them as rather unlikely to show adult content. The audio-based false positives show a screaming baby (j) and a video of a plane cockpit (i), in which we found audio noise similar to the one in many amateur porn clips. The two audio false negatives are simply overlayed with music (k,l). Finally, motion histograms produce false positives such as hand-filmed scenes with shaky cameras (m) and rhythmic motion such as the cheerleading lesson in (n). The false negatives in (o,p) show almost static scenes with little activity. For all these samples, joining in additional feature modalities helped improving recognition strongly, i.e. the displayed false positives as well as false negatives could be eliminated.

5. CONCLUSIONS

We have presented a quantitative study of pornography detection in video streams. While prior work has focused on using one or at most two feature modalities, our main contribution has been an assessment of four of the most widely used features. Our results indicate that a multi-modal approach seems vital for an accurate detection of adult content, and that even features that are rather inaccurate when applied individually (namely, motion and audio) can help increase accuracy in a multi-modal setting.

We have also studied a motion feature based on MPEG-motion compensation vectors: While a problem lies in the limited generalization of the approach between different input video codecs, results can be improved to some extent by adding image noise prior to motion estimation. Here, we plan to investigate the invariance of compressed-domain descriptors to codec-specific parameters further, employing other approaches such as a smoothing of the input signal.

6. ACKNOWLEDGEMENTS

This work has been funded by the EU Safer Internet Programme, Project iCOP (SI 2601002). We want to thank Christoph Lampert for his valuable comments on video compression.

7. REFERENCES

[1] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-Visual-Words Models for Adult Image Classification and Filtering. In *Proc. ICPR*, 2008.

[2] T. Endeshaw, J. Garcia, and A. Jakobsson. Classification of Indecent Video by Low Complexity Repetitive Motion Detection. In *Proc. AIPR Workshop*, 2008.

[3] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 88(2):303–338, 2010.

[4] M. M. Fleck, D. A. Forsyth, and C. Bregler. Finding Naked People. In *Proc. ECCV*, pages 593–602, 1996.

[5] U. Gargi and J. Yagnik. Solving the Label Resolution Problem in Supervised Video Content Classification. In *Proc. ACM MM*, pages 276–282, 2008.

[6] C. Jansohn, A. Ulges, and T. Breuel. Detecting Pornographic Video Content by Combining Image Features with Motion Information. In *Proc. ACM MM*, 2009.

[7] M. Jones and J. Rehg. Statistical Color Models with Application to Skin Detection. *IJCV*, 46(1):81–96, 2002.

[8] C.-Y.l Kim, O.-J. Kwon, W.-G. Kim, and S.-R. Choi. Automatic System for Filtering Obscene Video. In *Proc. ICACT*, pages 1435–1438, 2008.

[9] M. Kim and H. Kim. Automatic Extraction of Pornographic Contents using Radon Transform based Audio Features. In *Proc. Int. Workshop on Content-Based Multimedia Indexing*, pages 205–210, 2011.

[10] H. Lee, S. Lee, and T. Nam. Implementation of High Performance Objectionable Video Classification System. *ICACT*, pages 959–962, 2006.

[11] H.-T. Lin, C.-J. Lin, and R. Weng. A Note on Platt's Probabilistic Outputs for Support Vector Machines. *Mach. Learn.*, 68(3):267–276, 2007.

[12] Y. Liu, X. Wang, Y. Zhang, and S. Tang. Fusing Audio-Words with Visual Features for Pornographic Video Detection. In *Proc. TRUSTCOM'11*, pages 1488–1493, 2011.

[13] B. Logan. Mel Frequency Cepstral Coefficients for Music Modeling. In *Int. Symposium on Music Inf. Retrieval*, 2000.

[14] L. Rabiner and B.-H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, Inc., 1993.

[15] N. Rea, G. Lacey, C. Lambe, and R. Dahyot. Multimodal Periodicity Analysis for Illicit Content Detection in Videos. In *Proc. CVMP*, pages 106–114, 2006.

[16] H. Rowley, Y. Jing, and S. Baluja. Large Scale Image-Based Adult-Content Filtering. In *Int. Conf. Comp. Vis. Theory and Applications*, pages 290–296, February 2006.

[17] C. Snoek and M. Worring. Concept-based Video Retrieval. *Foundations and Trends in Information Retrieval*, 4(2):215–322, 2009.

[18] Xiaofeng T., L. Duan, C. Xu, Q. Tian, Hanqing L., J. Wang, and J.S. Jin. Periodicity Detection of Local Motion. In *Proc. ICME*, pages 650–653, 2005.

[19] A. Tourapis. Enhanced Predictive Zonal Search for Single and Multiple Frame Motion Estimation. In *Proc. SPIE Conf. Visual Communications and Image Processing*, pages 1069–1079, September 2002.

[20] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. A System that Learns to Tag Videos by Watching Youtube. In *Proc. ICVS*, pages 415–424, 2008.

[21] A. Ulges and A. Stahl. Automatic Detection of Child Pornography using Color Visual Words. In *Proc. ICME*, 2011.

[22] H. Zuo, O. Wu an W. Hu, and B. Xu. Recognition of Blue Movies by Fusion of Audio and Video. In *Proc. ICME*, pages 37–40, 2008.