Detecting Pornographic Video Content by Combining Image Features with Motion Information

Christian Jansohn
University of Kaiserslautern
D-67663 Kaiserslautern,
Germany
christian@jansohn.de

Adrian Ulges
German Research Center for
Artificial Intelligence (DFKI)
D-67663 Kaiserslautern,
Germany
adrian.ulges@dfki.de

Thomas M. Breuel
DFKI and University of
Kaiserslautern
D-67663 Kaiserslautern,
Germany
tmb@iupr.dfki.de

ABSTRACT

With the rise of large-scale digital video collections, the challenge of automatically detecting adult video content has gained significant impact with respect to applications such as content filtering or the detection of illegal material.

While most systems represent videos with keyframes and then apply techniques well-known for static images, we investigate motion as another discriminative clue for pornography detection. A framework is presented that combines conventional keyframe-based methods with a statistical analysis of MPEG-4 motion vectors. Two general approaches are followed to describe motion patterns, one based on the detection of periodic motion and one on motion histograms.

Our experiments on real-world web video data show that this combination with motion information improves the accuracy of pornography detection significantly (equal error is reduced from 9.9% to 6.0%). Comparing both motion descriptors, histograms outperform periodicity detection.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Filtering

General Terms

Algorithms, Measurement, Experimentation

Keywords

pornography detection; content-based video retrieval

1. INTRODUCTION

Increasing network bandwidth and storage capacity, video streaming, and web services like YouTube allow us to generate and share more digital video content than ever before (for example, 20 hours of video are uploaded to YouTube each minute). Correspondingly, new methods must be developed to cope with such huge loads of video data. One challenge in this context is the automatic detection of adult video content. Applications of this task include the filtering of personal video digest (e.g., for child protection) or digital forensics, where police forces are supported with the detection of illegal child pornographic content.

For the image domain, such technology has already been developed [2, 6, 11, 16] and is also applied in practice [11]. Regarding pornographic video material, the straightforward approach would be to extract representative *keyframes* and apply image classification techniques on them [7, 9].

In this paper, we demonstrate that better results can be achieved by enriching this standard approach with motion information. We compare keyframe-based methods (skin detection and bag-of-visual-words) with two motion analysis approaches (periodicity detection and motion histograms). Our evaluation is performed on real-world adult web videos and inoffensive content from the web portal YouTube. Results show that a significant improvement can be achieved by combining both information sources (image content and motion) in a late fusion step.

2. RELATED WORK

Most work regarding the detection of pornographic material has been done for the image domain. Forsyth et al. [5] proposed to detect skin regions in an image and match them with human bodies by applying geometric grouping rules. Wang et al. [16] presented a system that achieves a speed-up by a fast filtering of icons and graphs. Successive steps include the detection of skin areas and nearest-neighbor classification. Jones and Rehg focused on the detection of human skin by constructing RGB color histograms from a large database of skin and non-skin pixels [6], which allows to estimate the “skin probability” of a pixel based on its color. For adult image classification, simple features of the detected skin areas are fed to a neural network classifier (we will use a similar approach in our evaluation). Rowley et al. used Jones’ skin color histograms in a system installed in Google’s Safesearch [11]. Speed optimization was achieved by only extracting features in a small image area.

A different approach by Deselaers et al. [2] uses histograms of local image patches as a feature. Patches are extracted around difference-of-Gaussian interest points, described with their PCA transformation, and quantized with a codebook of patch categories (or *visual words*). A histogram over these patches is used as a feature vector for classification with a

MM’09, October 19–24, 2009, Beijing, China.
Copyright 2009 ACM 978-1-60558-608-3/09/10 ...$10.00.

Support Vector Machine (SVM). A similar approach will be included in our experiments.

Regarding the identification of offensive video material, fewer methods have been presented so far. Lee et al. [9] used a linear-discriminative classifier to combine two frame-based methods, one using on a skin probability map, the other color histograms. Kim et al. [7] used a shape description of skin areas in video frames. A manually defined color range is used for deciding whether a pixel belongs to a skin area. The area's shape is then described by normalized central moments and matched to samples in a database.

Also, other modalities have been employed for adult video content classification: Rea et al. [10] combined skin color estimation with the detection of periodic patterns in a video's audio signal (as pointed out, the method could similarly be applied to the motion signal). For periodicity detection, the surface of the lines through local maxima and minima in the signal's autocorrelation function is computed. Tong et al. [13] applied a similar method estimating the period of a signal to classify periodic motion patterns (we will include both approaches in our evaluation). Endeshaw et al. presented an approach that is entirely based on motion information [3]: repetitive motion patterns are detected by a spectral analysis using *periodograms*.

3. APPROACH

In the following, our framework for pornographic video detection is outlined. Given a video scene X, a *score* $P(o|X)$ is returned indicating the probability that X shows offensive material. Two general approaches to infer such scores are employed: image classification on keyframes (Section 3.1) and motion analysis (Section 3.2). Finally, a late fusion of the different methods is applied (Section 3.3).

3.1 Keyframe Analysis

For the keyframe-based classification, two methods are compared, one based on skin detection, the other on a patch-based bag-of-visual-words description.

Skin Detection (SKIN).

We evaluate a skin detection approach similar to Jones' method [6]: given an input image, a skin probability map is extracted using color histograms in RGB color space, and a binarization with a global threshold gives skin regions. Simple features including the average skin probability, the ratio of skin pixels, and the size of the largest skin region are used for classification with an SVM [1].

Bag-of-visual-words (BOVW).

The idea of this method is to describe images by histograms of local patches [12]. This description draws an analogy to the bag-of-words model from information retrieval, where documents (here, images) are represented by counts of words (here: *visual* words). Prior to feature extraction, a codebook of visual words is learned using a k-means clustering over image patches. Given a new image, its patches are matched to the closest visual word, and a histogram is constructed counting how often each visual word occurs in the image. This histogram is used as feature vector for classification with an SVM, which can be considered a cutting-edge approach for many recognition tasks such as concept detection [15] or object category recognition [4].

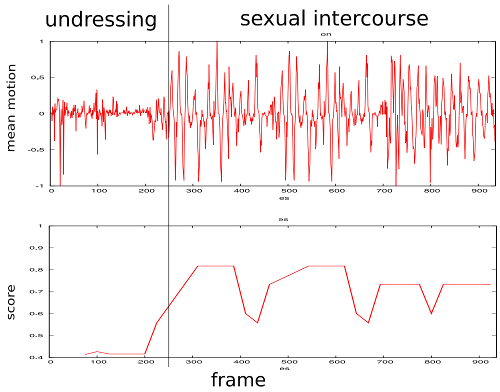

Figure 1: Periodicity detection (PERWIN): a video scene with its mean motion signal in x-direction $\bar{v}_x^t$ (top) and the classifier score (bottom). In the beginning of the video scene (left), clothes are taken off. Later, during sexual intercourse, periodic motion occurs, and scores indicate a higher probability for pornography.

In our implementation, overlapping patches of 14 × 14 pixels are regularly sampled at steps of 5 pixels, and the Discrete Cosine Transform (DCT) is used for patch description. To preserve color information, the DCT is applied to the luminance and both chrominance channels in YUV color space. 36 low-frequency components are used from the luma component and 21 for each chroma channel, giving a 78-dimensional descriptor per patch. The codebook consists of 2,000 visual words.

Scores $P(o|x_1), .., P(o|x_n)$ are obtained from a classification of keyframes $x_1, .., x_n$ and averaged over all keyframe scores (which has previously been demonstrated to give a high robustness with respect to noise in the input votes [8]).

3.2 Motion Analysis

Motion analysis is based on MPEG-4 motion vectors extracted by the XViD codec[1], such that each frame t is represented by a motion field $V_t = \{(x^{it}, v^{it}) | i = 1, .., n\}$. v^{it} denotes a 2D motion vector and x^{it} a macroblock position. The whole video is then represented as a sequence of motion fields $V_1, .., V_T$. In the following, we present two approaches for pornography detection based on an analysis of these fields.

Periodicity Detection (PER).

It seems reasonable to assume that sex scenes in video can be characterized by a periodic motion pattern. To capture this information, we use the *autocorrelation function* (ACF) as proposed by Rea et al. [10] and Tong et al. [13]. We combine both methods and apply them to the *mean motion* signals in x- and y-direction, $\bar{v}_x^t = \frac{1}{n} \sum_{i=1}^n v_x^{it}$ and $\bar{v}_y^t = \frac{1}{n} \sum_{i=1}^n v_y^{it}$. For both signals, we estimate the autocorrelation, and from it the periodicity as the mean distance between subsequent local maxima in the ACF [13]. Additionally, the variance of these distances is used. Also, the surface between the lines through the local maxima and minima is used as an additional feature that hints at the strength of periodicity [10]. This gives a six-dimensional vector, which is used for a decision tree classification.

[1] http://www.xvid.org

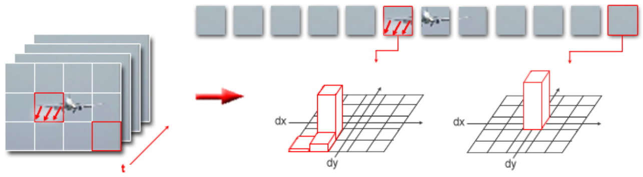

Figure 2: A visualization of motion histograms: frames are divided into 4×3 windows, and for each window a histogram over motion vectors is stored.

Sliding Window Periodicity (PERWIN).

Long offensive video tend to have only smaller parts where repetitive motion occurs. This motivates a slightly modified approach, where periodicity features are extracted over small sliding windows. A classification score is generated for each of these windows, and the final classification result is obtained by an averaging of these scores. The window size is three seconds, and one window is extracted per second.

A visualization of the approach is given in Figure 1: for a sex scene with two shots, the corresponding mean motion signal in x-direction is illustrated together with the score of our periodicity detector. In the first shot, almost no periodic motion occurs, which results in a low classification score. Later, sexual intercourse takes place and leads to a repetitive motion pattern, such that classification scores are higher.

Motion Histograms (MHIST).

This approach is based on motion histogram features by Ulges et al. [14], which describe *which* motion occurs as well as *where* it occurs. Each frame is divided into 4×3 regular blocks $B_1, .., B_{12}$, and for each of these blocks a motion histogram over all frames in the video is constructed (a visualization is given in Figure 2). The size of each histogram is 7×7 bins (for x- and y-direction). All motion vectors are clipped to $[-20, 20] \times [-20, 20]$. The final 588-dimensional feature vector is obtained by concatenating all block histograms $B_1, .., B_{12}$. This feature is fed to an SVM classifier.

3.3 Fusion

Finally, the classification scores $P_m(o|X)$ of the single methods m are combined in a late fusion step. A *weighted sum* fusion is used, where the influence of each method is represented by a corresponding weight $w_m \in [0, 1]$ learned from a validation set:

$$P(o|X) = \sum_m w_m P_m(o|X) \qquad (1)$$

4. EXPERIMENTS

To benchmark our system on real-world video data, we use offensive web videos and inoffensive YouTube videos. 932 Adult video clips were downloaded by a random crawl over unprotected pornographic websites, ranging from amateur videos to professional productions. YouTube was chosen for sampling a real-world mixture of non-pornographic content. 2,663 clips were downloaded using the YouTube API[2] for a variety of categories (animals, social events like concerts or dancing, nature, people, and sports). This gives a wide spectrum of inoffensive content, including many videos showing people (which is a harder and more realistic classification task than only focusing on nature themes).

Figure 3: Typical misclassifications of periodic motion detection indicate that repetitive motion patterns may be absent in adult video (left) but present in non-pornographic scenes, as in case of dancing (right).

All videos were scaled to a resolution of 320 × 240 pixels and converted using XViD for motion vector extraction. Pornographic video clips obtained by our crawler are typically of $10 - 30$ seconds length. As YouTube videos are usually longer, snippets of similar size as the adult videos were sampled randomly from these clips. Keyframes were extracted at regular steps of 50 frames, which gives $11,600$ keyframes for porn clips and $25,700$ for YouTube.

Performance evaluation was done using 5-fold cross-validation. The data was split into five equally sized sets, and three folds were used for training, one for validation (i.e., fitting the feature weights in Equation (1)), and one for testing. To avoid overfitting, we took care of the fact that some clips are from the same shoot, and placed them in the same set.

Quantitative results in terms of equal error rate and ROC curves (averaged over 5 runs) are given in Figure 4. We first compare the individual approaches. When focussing on keyframe-based methods (SKIN, BOVW), it can be seen that the bag-of-visual-words method outperforms skin detection (which confirms earlier results by Deselaers [2]). Regarding the motion-based approaches, periodicity detection with a sliding window works better than on the whole video. This can be explained by the fact that a sliding window approach is more robust to small breaks in the motion pattern, which frequently occur even in relatively short sequences of less than 30 seconds length.

Yet, even this improved periodicity-based approach does not reach the performance of motion histograms, which work significantly better and almost reach the performance of the best keyframe-based approach (equal error rate: 12.52%). Obviously, motion histograms are more appropriate for capturing motion patterns that occur only in small areas of the frame. Also, it seems that many discriminative motion patterns for adult content are not necessarily strictly periodic – not all pornographic videos show repetitive motion, and there are also inoffensive videos with strong periodic motion patterns, like dancing in music videos (see Figure 3).

Next, the system is evaluated when combining several methods in a late fusion. Again, see Figure 4 for quantitative results. It can be seen that a strong classification performance is achieved by fusing bag-of-visual-words with motion histograms, which outperforms both single descriptors. Compared with the best system using a single modality (DCT, equal error 9.88%), a significant performance increase is achieved (equal error 6.04%). A fusion of all feature types did not give any further improvements compared to this BOVW+MHIST system.

[2] www.youtube.com/dev

Feature	μ-EER	w_1	w_2
BOVW	**9.9** ± 0.45	-	-
SKIN	18.35 ± 0.80	-	-
MHIST	**12.52** ± 1.02	-	-
PER	37.85 ± 0.79	-	-
PERWIN	28.33 ± 0.41	-	-
BOVW+MHIST	**6.04** ± 0.52	0.54	0.46
BOVW+PERWIN	8.56 ± 0.42	0.45	0.55
SKIN+MHIST	10.97 ± 1.09	0.59	0.41
SKIN+PERWIN	17.43± 0.87	0.56	0.44

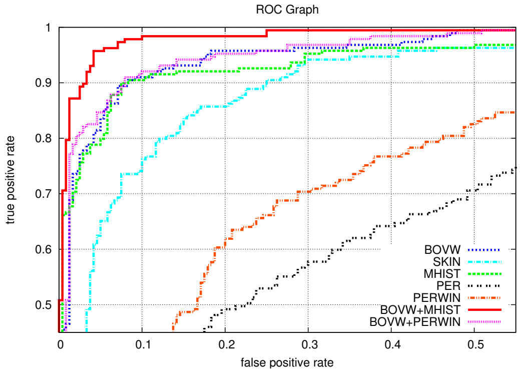

Figure 4: Quantitative results in terms of equal error rate (EER, top) and ROC curves (bottom). The best error rate for a single feature is achieved by the bag-of-visual-words approach (BOVW). Motion histograms (MHIST) outperform periodicity detection (PER, PERWIN). A combination of visual words and motion histograms gives a significant improvement (error is reduced from 9.9% to 6%).

5. DISCUSSION

In this paper, we have addressed the automatic detection of pornographic content in video databases, a problem with practical applications in content filtering and digital forensics. Our key contribution lies in the fact that we evaluate both image features and motion information as discriminative clues for the detection of pornographic material. To the best of our knowledge, the study presented in this paper is the first one that compares both feature modalities. Particularly, we show that significant improvements can be achieved by a combination of image features and motion in a simple late fusion step.

Correspondingly, one future direction along the proposed line of research is the use of audio as a third modality. Other options might be a hierarchical approach to speed up the system (for example, a fast skin-based approach or motion analysis can be used to rule out simple cases quickly, and ambiguous material is examined by a more accurate but cost-intensive patch-based model).

Finally, we also believe that repetitive motion detection deserves more investigation. Our results with approaches from the literature [10, 13] did not validate improvements over motion histograms or image features, which is to some extent caused by a lack of robustness with respect to breaks in the motion pattern. In contrast, results by Endeshaw et al. [3] following a slightly different approach indicate that repetitive motion detection *can* be helpful for long-term motion patterns, such that repetitive motion might be investigated further[3].

6. REFERENCES

[1] C. Burges. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Min. Knowl. Discov.*, 2(2):121–167, 1998.

[2] T. Deselaers, L. Pimenidis, and H. Ney. Bag-of-Visual-Words Models for Adult Image Classification and Filtering. In *ICPR*, pages 1–4, December 2008.

[3] T. Endeshaw, J. Garcia, and A. Jakobsson. Classification of Indecent Video by Low Complexity Repetitive Motion Detection. In *Proc. 37th Applied Imagery Pattern Recognition Workshop*, 2008.

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

[5] M. Fleck, D. Forsyth, and C. Bregler. Finding Naked People. In *ECCV (2)*, pages 593–602, 1996.

[6] M. Jones and J. Rehg. Statistical Color Models with Application to Skin Detection. *Int. Journal of Comp. Vision*, 46(1):81–96, 2002.

[7] C.-Y. Kim, O.-J. Kwon, W.-G. Kim, and S.-R. Choi. Automatic System for Filtering Obscene Video. In *ICACT*, volume 2, pages 1435–1438, 2008.

[8] J. Kittler, M. Hatef, R. Duin, and J. Matas. On Combining Classifiers. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20(3):226–239, 1998.

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

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

[11] H. Rowley, Y. Jing, and S. Baluja. Large Scale Image-Based Adult-Content Filtering. In *VISAPP (1)*, pages 290–296, 2006.

[12] J. Sivic and A. Zisserman. Video Google: A Text Retrieval Approach to Object Matching in Videos. In *ICCV*, pages 1470–1477, 2003.

[13] X. Tong, L. Duan, C. Xu, Q. Tian, Hanqing L., J. Wang, and J. Jin. Periodicity Detection of Local Motion. *ICME*, pages 650–653, 2005.

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

[15] K. van de Sande, T. Gevers, and C. Snoek. A Comparison of Color Features for Visual Concept Classification. In *CIVR*, pages 141–150, 2008.

[16] J. Wang, G. Wiederhold, and O. Firschein. System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms. In *IDMS*, pages 20–30, 1997.

[3] This work was supported by the EU Safer Internet Programme, project FIVES (SIP-2008-TP-131801), and by the German Research Foundation (DFG), project MOONVID (BR 2517/1-1).