RECOGNITION OF BLUE MOVIES BY FUSION OF AUDIO AND VIDEO

Haiqiang Zuo, Ou Wu, Weiming Hu, Bo Xu

National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, China

ABSTRACT

Along with the explosive growth of the Internet, comes the proliferation of pornography. Compared with the pornographic texts and images, blue movies can do much harm to children, due to the greater realism and voyeurism of blue movies. In this paper, a framework for recognizing blue movies by fusing the audio and video information is described. A one-class Gaussian mixture model (GMM) is used to recognize porno-sounds. A generalized contour-based pornographic image recognition algorithm is used to detect pornographic image frames of a video shot. Then a fusion algorithm based on the Bayes theory is employed to combine the recognition results from audio and video. Experimental results demonstrate that our framework which exploits both audio and video modalities is more robust and achieves better performance than one which uses either one alone.

***Index Terms*—**Blue movies, pornography, multimodal fusion

1. INTRODUCTION

In the last few years, the explosive growth of the Internet, World Wide Web, is reshaping the nature and environment of pornography at an accelerated pace. The Internet offers nearly free access to pornography uninhibited by previous barriers of time and space. Pornography has attracted users on line and it has given many people a reason to spend time surfing the Internet [1]. In recent years a substantial literature on web pornography filtering has developed. Approaches mainly focus on detecting pornographic texts and images [2-5]. Compared with the pornographic texts and images, blue movies can do much harm to children, due to the greater realism and voyeurism of blue movies. However, little work has been done to recognize blue movies on the web. Technologies to recognize and block blue movies should be fully developed.

In movies, the image frames and the audio waveform are closely related and provide the same kind of information. Though using audio or video recognition system respectively may provide the expected classification result, it is prone to failure for each of them. The result depends on not only the goodness of the classifiers but also the characteristics of the scene such as the video illumination and audio noise. The more reliable approach is to construct a framework that can fuse the available information in audio tracks and video frames of a movie.

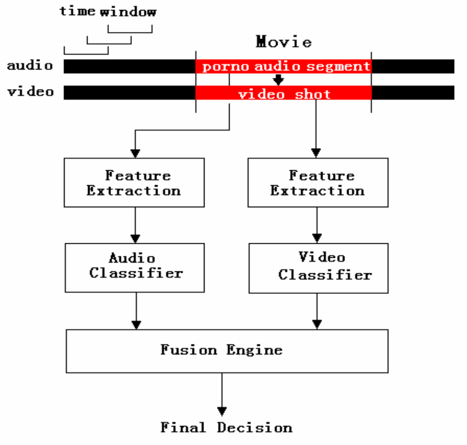


Figure 1. Overview of our framework.

Figure 1 shows an overview of our framework. Because audio analysis usually needs less computation time than video analysis, we scan audio tracks first. We extract the desired audio features from a short (1-5 seconds) window, and a likelihood value is computed by the audio classifier. If the likelihood value is greater than 0.5, this window is marked as "porno". The window goes forward with 50% overlap. A series of continuous porno windows form a porno audio segment. When a porno audio segment is detected, the corresponding video shot is derived. The contour-based features are extracted to recognize pornographic image frames of the video shot. The results of the audio classifier and the video classifier are incorporated into our fusion engine. The algorithm stops, if the movie is classified as a blue one, and otherwise the process repeats until the end of the movie. If no porno audio segment is detected, the video classifier is used alone.

The remainder of this paper is organized as follows: Section 2, 3 introduce the porno-sounds recognition algorithm and the pornographic image frames recognition algorithm respectively. Section 4 describes the fusion algorithm. Section 5 demonstrates experimental results. Section 6 summarizes this paper.

978-1-4244-2571-6/08/$25.00 ©2008 IEEE

2. PORNO-SOUNDS RECOGNITION

In blue movies, the audio and video signals are closely related. The specific sounds in a blue movie are often the women moan or scream(we call them porno-sounds in this paper).

2.1. Audio Feature Extraction

One of the most popular set of features used to parameterize the waveform is the Mel-Frequency Cepstral Coefficients (MFCC). MFCC is a sub-band energy feature in mel-scale, which gives a more accurate simulation of human auditory system. This parameterization has been demonstrated to provide good representations of a speech signal allowing for better discrimination than temporal or frequency based features alone [6]. In our framework, the audio waveform is transformed into a sequence of 13-dimensional feature vectors (12 MFCC coefficients plus the energy term). Each sound consists of a mass of points in MFCC feature space, and for the sake of simpleness and computational cost, each sound here is represented by the mean of the sequence of MFCC vectors, and thus each sound is reduced to a single point in MFCC feature space.

2.2. Audio Classification

To be able to recognize porno-sounds, a one-class Gaussian mixture model (GMM) is built. The reason to choose one-class classifier is that only one of the classes, the porno-sound class, can be sampled well while the other class, the normal sound class, is difficult to collect, because the normal sounds are so abundant that a good sampling of them is not possible. In one-class classification we are always dealing with a two-class classification problem, where each of the two classes has a special meaning. The two classes are called the target and the outlier class respectively [7]. In this case, the porno-sounds are the targets, and the normal sounds are outliers. The target class is modeled as a Gaussian mixture distribution:

$$p(x;\mu_k,\Sigma_k,\pi_k)=\sum_{k=1}^{m}\pi_k p_k(x),\quad \pi_k\geq 0,\quad \sum_{k=1}^{m}\pi_k=1, \tag{1}$$

$$p_k(x)=\frac{1}{\sqrt{(2\pi)^{d/2}\left|\Sigma_k\right|^{1/2}}}\exp\{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)\}, \tag{2}$$

where m is the number of mixture components, p_k is the norm distribution density with the mean vector μ_k and covariance matrix Σ_k, π_k is the weight of k-th mixture, d is the vector dimension. An appropriate value for the number of mixtures m can be determined by the experimental results which the best performance reaches. The expectation-maximization (EM) algorithm [8] is used to calculate GMM parameters. Given the number of mixtures m and the training samples $\{x_i,\ i=1...N\}$ the algorithm finds the maximum-likelihood estimates (MLE) of the all the mixture parameters. The tradeoff between the target acceptance rate and the outlier rejection rate is present in all one-class classification methods. It is possible to set a threshold θ_d, that a prespecified target acceptance rate (0.95 in our case to capture most of the training sounds) on the training set is obtained. Let d_m be the Mahalanobis distance from the test point to the mean of the learned GMM which minimizes this distance. If $d_m<\theta_d$ the sound is classified as a target, and otherwise the sound is classified as an outlier. For the later fusion purpose, we define a likelihood value L_t by

$$L_t=\begin{cases}1 & \text{if } d_m<\dfrac{\theta_d}{2}\\[2ex] \dfrac{\theta_d}{2d_m} & \text{otherwise.}\end{cases} \tag{3}$$

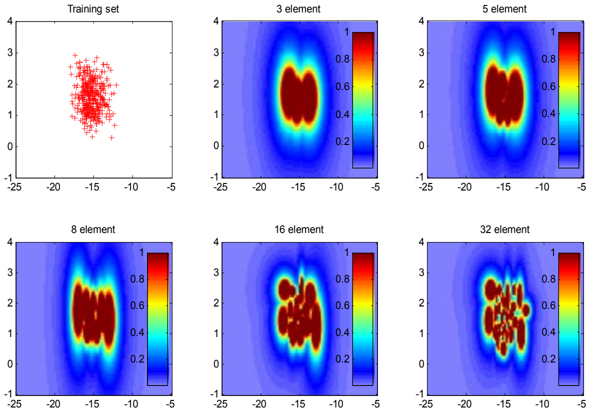


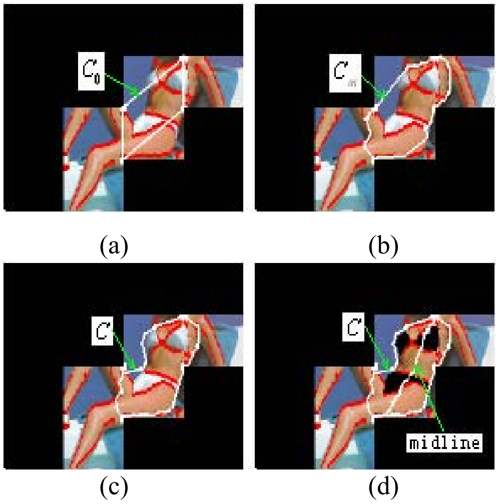
Figure 2. The training set (top left) and the likelihood value regions of different mixture numbers.

Thus the points in the feature space are mapped into an interval (0, 1]. The points near the means of GMM have high values, and those far from the means of GMM have low values, and when a point is located on the class boundary the likelihood value has the output 0.5. This likelihood is essentially equivalent to the log-likelihood but is fast and convenient to compute.

Figure 2 shows our training set and the likelihood value regions generated by the GMM trained on the 592 porno-sounds of different mixture numbers. For illustration, only the fist two features (the energy term and first MFCC coefficient) are used.

3. PORNOGRAPHIC IMAGE FRAMES RECOGNITION

To recognize pornographic image frames of a video shot, the contour-based pornographic image recognition algorithm [4] (our previous work) is used, and is extended to produce a generalized version in this work. In our previous algorithm, the image plane is partitioned into 4×4 rectangular blocks. In this generalized algorithm, the image plane can be partitioned into any number of rectangular blocks. The algorithm uses effective human shape constraints to distinguish pornographic image frames from normal image frames rich in skin colors.



(a) (b)

(c) (d)

Figure 3. The contour of a body trunk [4]. (a) The initial curve. (b) The middle stage curve. (c) The refined curve. (d) The midline and non-skin regions.

Figure 3 shows the process of extracting the contour of a body trunk. Fist, the skin edges are detected in the ROIs (region of interest), and the initial closed curve C_0 is obtained by connecting all the POIs (point of interest), as shown in Figure 3a. Second, the initial closed curve C_0 and the skin edges outside curve C_0 are merged to form a middle stage curve C_m, as shown in Figure 3b. Third, all the non-skin pixels on curve C_m are adjusted to the near skin edges and the refined curve C is obtained, as shown in Figure 3c. At the final stage, the midline of refined curve C and non-skin regions within curve C are detected, as shown in Figure 3d. See [4] for more details. The image features are then extracted from the derived contour, and a nearest-center classifier is employed to distinguish pornographic image frames from normal image frames.

4. FUSION OF AUDIO AND VIDEO

In this section, our fusion method is described for fusing the information obtained from image frames and the information obtained from audio tracks. In blue movies, the image frames and the audio waveform are closely related and provide the same theme — obscenity. Based on this observation, we assume the following priori knowledge:

- Almost all image frames in a video shot are pornographic in a blue movie or normal in a normal movie.
- The information in audio tracks provides prior knowledge of image frames.

Then, the problem of classifying a video shot becomes a problem of classifying a set of image frames, where the result of classifying audio tracks in a movie is used as the prior knowledge. We define two statistical features of the image frame classifier:

- the probability p_1 that a normal image frame is mistakenly classified as a pornographic one, and
- the probability p_2 that a pornographic image frame is mistakenly classified as a normal one.

The probabilities p_1 and p_2 can be estimated statistically by counting the number of image frames mistakenly classified by the pornographic image recognition algorithm in a large set of known image frames. Suppose that there are N image frames in a video shot and the result of our image frame classifier is that N_1 image frames are classified as pornographic and N_2 ones as normal (N_2=N-N_1). Let r = (N_1 pornographic image frames, N_2 normal image frames). Let S represent the event that all the N image frames are pornographic and $\neg S$ the event that all the N image frames are normal. Then, the equations below are obtained:

$$p(r \mid S) = (1-p_2)^{N_1}(p_2)^{N_2}, \tag{4}$$

$$p(r \mid \neg S) = (p_1)^{N_1}(1-p_1)^{N_2}. \tag{5}$$

According to the Bayes rule, the following equations are obtained:

$$p(S \mid r) = \frac{p(r \mid S) \times p(S)}{p(r)}, \tag{6}$$

$$p(\neg S \mid r) = \frac{p(r \mid \neg S) \times p(\neg S)}{p(r)}. \tag{7}$$

We introduce a decision factor f, which is the ratio of the two posterior probabilities in (6) and (7):

$$f = \frac{p(S|r)}{p(\neg S|r)} = \frac{p(r|S) \times p(S)}{p(r|\neg S) \times p(\neg S)} = \frac{(1-p_2)^{N_1}(p_2)^{N_2}}{(p_1)^{N_1}(1-p_1)^{N_2}} \times \frac{p(S)}{p(\neg S)}. \tag{8}$$

If $f \geqslant 1$, the movie is classified as a blue one.

The remaining problem is to evaluate the priori probabilities $p(S)$ and $p(\neg S)$. As audio track provides prior knowledge of image frames, $p(S)$ is replaced with the likelihood L_t introduced in Section 2.2, and $p(\neg S)$ is set equal to 1- L_t. Then,

$$f = \frac{(1-p_2)^{N_1}(p_2)^{N_2}}{(p_1)^{N_1}(1-p_1)^{N_2}} \times \frac{L_t}{1-L_t+\varepsilon}, \tag{9}$$

where ε >0 is a small enough number. To avoid the error of dividing by zero, the log version of equation (9) is used.

$$\begin{aligned} F &= [N_1 \log(1-p_2) + N_2 \log(p_2) + \log(L_t)] \\ &\quad -[N_1 \log(p_1) + N_2 \log(1-p_1) + \log(1-L_t+\varepsilon)]. \end{aligned} \tag{10}$$

If $F \geqslant 0$, the movie is classified as a blue one.

5. EXPERIMENTAL RESULTS

5.1. Porno-Sounds Recognition

To evaluate the performance of our one-class GMM porno-sounds recognition algorithm, 1,412 sounds were manually collected and labeled, including 592 porno-sounds as the training set, and 820 sounds (including 268 porno-sounds and 552 normal sounds) as the test set. A wide variety of normal sounds are represented from animals, birds, nature, musical instruments, speech, vehicles, sports and recreation. The sounds vary in duration from less than a second to about 10 seconds. All audio streams are in the 22,050 Hz, 16-bit and mono channel format, and are divided into frames of 16 ms with 50% overlap for feature extraction.

Table 1. Results of porno-sounds recognition algorithm.

Gauss elements	Recall	Precision	F_1
3	94.4%	86.7%	90.4%
5	95.7%	84.8%	89.9%
8	95.9%	86.6%	91.0%
16	96.4%	88.3%	92.2%
32	94.4%	84.9%	89.4%

Table 1 shows the performance of our one-class GMM porno-sounds recognition algorithm with different number of mixture components. It can be seen that when the number of mixture components *m* is set to 16, the algorithm reaches the best performance.

5.2. Pornographic Image Frames Recognition

To evaluate the performance of our generalized contour-based pornographic image recognition algorithm, 53,554 pornographic image frames and 83,548 normal image frames were manually extracted and labeled from a set of movies.

Table 2. Results of pornographic image frames recognition algorithm.

Recall	**Precision**	F_1
86.7%	78.8%	82.6%

The results are presented in table2. The error rates are relatively high compared with the results of [4] because of the lower image quality and the more complexity of video scene than web images. The probabilities p_1 and p_2 introduced in Section 4 are estimated to be 0.150 and 0.133 respectively.

5.3. The Fusion of Audio and Video

To evaluate the performance of our framework, 352 blue movies and 537 normal movies were downloaded from the Internet. Comparison experimental results of using different modalities to recognize blue movies on the same data set are shown in Table 3.

When using audio modality alone, we count the number of porno audio segments detected by our one-class GMM porno-sounds recognition algorithm in audio tracks, and if a sufficient number of porno audio segments are detected, the movie is labeled as blue. Similarly, when using video modality alone, we count the number of detected pornographic image frames and if the number of pornographic image frames exceeds a threshold, the movie is labeled as blue. The experimental results show that our fusion method which exploits both audio and video modalities has 8.9% - 12.8% improvements on the F_1 measure than using the audio or video modality alone.

Table 3. Comparison results of different modalities.

Modalities	**Recall**	**Precision**	F_1
Audio only	90.1%	82.8%	86.3%
Video only	84.4%	80.5%	82.4%
Audio and video	98.3%	92.3%	95.2%

6. CONCLUSIONS

In this paper, we propose a framework which exploits both audio and video modalities to recognize blue movies. Experimental results demonstrate that our framework is more robust and effective than one which uses either modality alone. Each modality may compensate for weaknesses of the other one.

7. ACKNOWLEDGMENT

This work is partly supported by NSFC (Grant No.60672040, 60705003) and the National 863 High-Tech R&D Program of China (Grant No. 2006AA01Z453).

8. REFERENCES

[1] J. Coopersmith, "Pornography, videotape and the Internet," *IEEE Technology and Society Magazine*, pp.27-34, 2000.

[2] J. Z. Wang, J. Li, G. Wiederhold and O. Firschein, "System for screening objectionable images," *Computer Communication Journal*, pp.1355-1360, 1998.

[3] W.H. Ho, P.A. Watters, "Identifying and Blocking Pornographic Content," *21st International Conference on Data Engineering Workshops*, pp. 1181-1188, 2005.

[4] J. Yang, Z. Fu, T. Tan, and W. Hu, "A Novel Approach to Detecting Adult Images," *Proc. Int'l Conf. Pattern Recognition*, pp.479-482, Aug. 2004.

[5] J. Ruiz-del-Solar, V. Castaneda, R. Verschae, R. Baeza-Yates, F. Ortiz, "Characterizing Objectionable Image Content (Pornography and Nude Images) of Specific Web Segments: Chile as a Case Study," *Third Latin American Web Congress*, pp. 269-278, 2005.

[6] A. H. Sadka, "Visnet: NoE on Networked Audiovisual Media Technologies," *Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, April 2004.

[7] Tax D.M.J. *DDtools, the Data Description Toolbox for Matlab*. version 1.6.0, 2007.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum-likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society (series B)*, pp.1-38, 1977.