Erotic Audio Recognition Using Heterogeneous Ensemble Classifiers

Ziqiang Shi, Tieran Zheng, Jiqing Han, and Boyang Gao

***Abstract*—In this paper, we present a novel approach to the task of automatic adult video detection based on feature porno-sounds recognition using multiple feature vectors and an ensemble of binary classifiers. In this framework, firstly the audio track is extracted and segmented into equal segments from an unknown video. Then each segment is split into sequences of voiced and unvoiced unequal fragments. Multiple feature vectors are extracted for each voiced fragments. Components classifiers including Support Vector Machine (SVM) are trained based on such feature vectors. At the classification, the outputs provided by the component classifiers are combined through fusion rules to form a final output of the ensemble. According to the classification results of these fragments, each segment is classified as erotic or natural. Finally, based on these segments results, the audio track or further the unknown video is labeled "erotic" or "natural". Online and off-line experiments show that the proposed approach yields high performance than using single classifiers.**

***Index Terms*—Ensemble classifiers, SVM, artificial neural network (ANN), erotic audio, voiced segment**

I. Introduction

Nowadays the Internet provides a huge collection of multimedia documents, but it also contains large amounts of pornographic materials, such as adult video sequences, which are unsuitable for children and unsuspecting adults. Thus effective pornographic video detection is important for preventing social-cultural problems.

Previous work in this field has mainly focused on processing text or visual data, e.g. [1], [2]. Due to the fact that adult videos are always accompany particular human sounds, such as groans of females which are distinct different from natural sound, a method from the feature porno-sounds recognition point of view is proposed to detect adult video sequences in [3], [4].

This problem of erotic sound detection belongs to the more general field of audio event detection (AED) or audio classification. Past work always does audio classification at the unit of one second or one frame [5], [6]. This rigid unit is not suitable for our problem. Fig. 1 shows the spectrogram of a typical erotic audio clip. It can be seen that different erotic sounds always have different durations. It is more reasonable to set one complete groan or scream as the basic unit for recognition and classification in our task. In this paper, we set the continuous voiced fragment as the smallest unit for recognition. Means and variances of mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), pitch, and short-time energy computed from each fragment are used as features.

For classifiers, we can combine a number of member classifiers to yield a more accurate recognition rates [7-9]. This approach is known as the ensemble of classifiers. The features are used to train different types of classifiers: SVM, Gaussian Mixture Models (GMM) based classifier and ANN. In this work, the outputs of the component classifiers are fused through two combination methods, including majority voting and fusion classifier.

The organization of the paper is as follows. Section II covers an overview of the voiced fragments classification based erotic audio recognition system. Section III presents the features extracted from voiced fragments. The member classifiers and ensemble strategies are presented in section IV. Section V introduces the rejection and verification modules of the system. Section VI shows the experimental set-up and results. Finally conclusions are drawn in section VII.

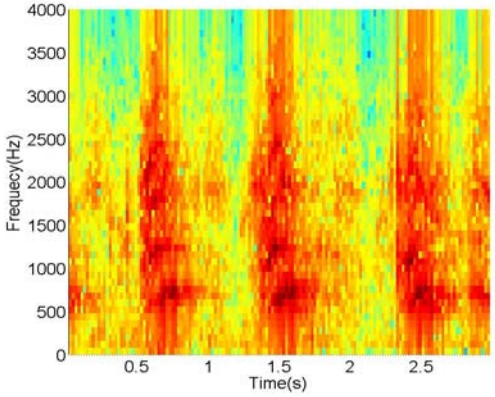


Fig. 1. Spectrogram of a typical erotic audio clip

II. Erotic Audio Recognition System

The erotic audio recognition system proposed in this paper is composed of three main phases: feature extraction, classification and decision. First, the audio signal is segmented into equal segments. Then we split each segment into sequences of voiced and unvoiced unequal fragments. Features are extracted from these voiced fragments.

For classification, ensemble classifier is employed for its robustness and effectiveness. In the training mode, component classifiers are trained using the extracted features. In the classification mode, for an unknown audio track, the

Manuscript received September 12, 2012; revised October. This work was partly supported by the National Natural Science Foundation of China under grant No. 91120303 and No. 61071181.

Ziqiang Shi, Jiqing Han, Tieran Zheng are with the School of Computer Science and Technology, Harbin Institute of Technology, No. 92, West Da-Zhi Street, Harbin, Heilongjiang, China (Tel: +86-451-86417981, e-mail: zqshi, jqha@hit.edu.cn n, zhengtieran@hit.edu.cn).

Boyang Gao is with the Department of Mathematics and Computer Science Ecole Centrale de Lyon, Lyon, France (e-mail: gaoboyang@gmail.com).

features are fed into the trained component classifiers. The outputs of the classifiers are then fused through majority voting or fusion classifier. Based on the results, labels of "pornographic" or "natural" are assigned to the voiced fragments in the 3-sec segment. If the number of erotic fragments is large than a threshold, then this segment is labeled "pornographic", else it is labeled "natural". Then the audio signal is sent to a verification module and the long term similarity measures are obtained. If the number of erotic segments in an audio track and the long term similarity measures are larger than thresholds, then raise the alarm. Fig. 2 illustrates such a process.

In the next sections, the important components of the system are described, such as voiced fragments search and feature extraction, component classifiers and ensemble methods, rejection and verification modules.

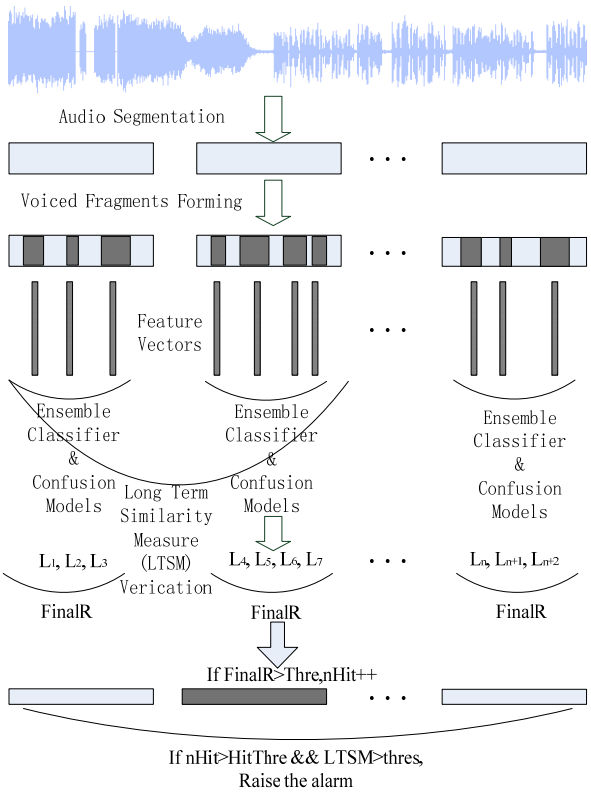


Fig. 2. The proposed erotic audio recognition system

III. Feature Extraction

A. Basic Feature Extraction

At the first step, the audio stream is windowed into a sequence of short-term frames (20 ms long) with 10-ms overlap. Four basic features are extracted per frame. Briefly, we used two percept features, namely the short-time energy and the pitch. Short-time energy is the energy of a frame and is calculated as the sum of squared amplitudes within a frame. In this paper, we used a simple pitch detection algorithm based on detection peak of the normalized autocorrelation function. Furthermore, we used the MFCCs and LPCCs. MFCCs are a perceptually motivated representation of the coarse shape of the signal spectrum. We used 12 MFCC coefficients calculated from the outputs of a 20-channle filter bank. LPCCs were obtained using the autocorrelation method. The number of LPCC coefficients extracted was 10. For notation purpose, let $\mathrm{L}=\{M_t, L_t, P_t, E_t\}$, $t=1\cdots T$ be the resulting feature sequence, where T is the number of frames, and M_t, L_t, P_t, E_t stand for the t-th MFCCs, LPCCs, pitch and energy respectively.

B. Voiced Fragments Search and Feature Extraction

The first stage of the system is to segment the unknown audio track into equal segments. In this paper, each segment has a length of 3 seconds, that's to say each segment contains 299 20-ms frames.

In the second stage, we search for voiced fragments in every segment. Voiced fragment means a short fragment that almost all frames in it have pitches. The details of the proposed voiced fragment search algorithm are illustrated below.

procedure $Vf \leftarrow \mathrm{Vfs}(S)$

//input: S, the 3-sec segment.

//output: Vf, the set of voiced fragments found in S

Begin

1) Extract the feature sequence of S, which is denoted by $\mathrm{L}=\{M_t, L_t, P_t, E_t\}, t=1...T$, set s =1, $Vf \leftarrow \varnothing$.
2) Start from the frame s, find the first index i that P_i is larger than 10, this is the start frame of a new voiced fragment. If all P_t are less than 10, goto End.
3) Check the frames following the frame i, find the longest fragment F satisfy that there is no continuous 5 frames that all P s are less than 10, then $Vf \leftarrow F$ and set s the index of the first frame following F in S.

Repeat 2 and 3 until reach the end of L.End

Now, for each measure quantity (MFCCs, LPCCs, pitch, and short-time energy), mean and variance computed from respective voiced fragments is used as a feature. That's to say each feature vector corresponding to a voiced fragment. The final feature vector concatenates all these features including mean and variance of MFCCs (24), mean and variance of LPCCs (20), mean and variance of pitch (2), mean and variance of energy (2) into a 48-dimensional feature vector.

IV. Ensemble of Classifiers

A. Component Classifiers

In this work we use the following machine learning methods as component classifiers: SVM, GMM and ANN. SVM is primarily a method that performs classifications by constructing hyper-planes in a higher dimensional that separates cases of different class labels. In this work we have used two classes SVM.

GMM-based classifiers have shown increased attention in many pattern recognition tasks. In this paper, we use 128 Gaussians in GMM and the training is performed independently for each class. The neural network is composed of 48 neurons in the input layer, one neuron in the

output layer, and four hidden layers with 96, 48, 24, 6 neurons respectively.

B. Ensemble Method

We denote component binary classifier by $\Phi(\bullet)$. The individual decisions of all classifiers in the ensemble are combined [12]:

$$\Theta(x) = \begin{cases} 1 & \text{if } g(\Phi_1(x), \Phi_2(x), ..., \Phi_m(x)) > 0, \\ -1 & \text{otherwise,} \end{cases} \quad (1)$$

where $\Theta(\bullet)$ is an ensemble classifier; $g(\bullet,...,\bullet)$ is a combination function that combines the outputs of all component classifiers $\Phi_i(\bullet)$. The effectiveness of $g(\bullet,...,\bullet)$ decide the effectiveness of the ensemble classifier $\Theta(\bullet)$. In this paper, we use two different $g(\bullet,...,\bullet)$:

- Majority Voting (MV):
  Majority voting is one of the simplest and most intuitive ensemble combination techniques. It is a simple decision rule where only the class labels are taken into account and the one with more votes wins. In MV, all the classifiers in the ensemble are equally weighted.
- Fusion Classifier (FC):
  Let $\Phi_1(x)$, $\Phi_2(x)$, …, $\Phi_m(x)$ be m different binary classifier and their outputs forms a new vector. These new vectors are used to train a new classifier which is used to fuse the outputs of all the component classifiers. In this work, a SVM is used as the fusion classifier.

In Section VI, the results of the ensemble methods are evaluated relative to the conventional classification approaches that use single classifier.

V. Rejection and Verification Modules

In order to increase the efficiency of the system, we implement a rejection module based on confusion model and a verification module based on global similarity in the system.

The objective of rejection and verification is to improve the classification accuracy by redirecting samples with high uncertainty to a specialized classification stage. In our problem, few natural sounds share some perceptual feature with erotic sounds, such as soprano, stringed music, and so on. These similar sounds form the ambiguous set, but this set is not explicit. In order to obtain this set, we run the system without rejection module online for 48 hours, and then pick out the non-target audio tracks that the system alarmed. Confusion models are trained on this ambiguous set against the original target set. Reject classifiers are used together with recognition classifier to the features extracted from the voiced fragments. Experiments show that the rejection module improves the performance of the system.

From observation and experiments, the erotic audio always shows a quasi-periodicity in a long time section. For a long time section (e.g. 15 sec), after the voiced fragments searching and feature extraction, we find the average cosine similarity of all pairs of features in this section. This is called the long term similarity measure (LTSM) of this long time section of the audio. Statistics show that the LTSMs of erotic audio concentrate in a short interval near one, while the LTSM of nature audio has a flat distribution. So a predefined LTSM threshold will filter out part of the nature audio. In this paper, the LTSM threshold is set to be 0.65.

VI. Experiments

A number of uninterrupted audio streams have been extracted from more than 50 adult movies of various kinds. These audio streams were manually segmented and the segments with typical erotic sounds were labeled "pornographic". The "nature" data were recorded from the Internet and manually selected. In total, almost 4 hours of "pornographic" and 20 hours of "nature" audio have been used for training the component classifiers and the fusion classifier.

Three measures are used for describing the performance of the system:

- Precision: The proportion of audio data that was classified as pornographic and was indeed pornographic.
- Recall: The proportion of pornographic data, which was correctly classified as pornographic.
- False Accept Rate: The proportion of audio data that was classified as pornographic and indeed was not pornographic.

A. Off-line Experiments and Results

For the off-line experiment, 1152 audio files were extracted from respective video files of various formats which were recorded randomly online. These files were independent from training files. This test set was manually labeled, 106 files were pornographic ones. The experiments compare the erotic audio recognition system with individual classifiers (EAR-SVM/GMM/ANN), ensemble of classifiers with majority voting (EAR-MV) and fusion classifier (EAR-FC).

In Table I, false accept rate and recall are presented for all systems. The ensemble classifiers achieved better performance compared to most of the single component classifier results. In the case of EAR-SVM the false accept rate was improved in more than 10%; more than 4% and 6% for the EAR-ANN, 3% and 10% about for the EAR-GMM respectively on recall and false accept. For ensemble methods, EAR-FC has a better performance than EAR-MV.

TABLE I: System Performance in off-Line Experiments

Approach	Recall (%)	False Accept Rate (%)
EAR-SVM	84.9	19.3
EAR-GMM	70.7	19.0
EAR-ANN	69.8	14.9
EAR-MV	75.4	8.9
EAR-FC	73.5	2.7

As far as the false accept rate of the method is concerned, the best false accept rate is 2.7%, while the recall is 73.5%. In other words, for every 10 detected videos almost all are indeed adult videos, while more than 70% of the adult videos are detected. The reason for not high recall can be that the training data did not cover all kinds of the particular sound

scenes in adult videos and the classifier cannot recognize these scenes.

B. Online Experiments and Results

Due to the fact that it is impossible to check tens of thousands Internet video files, in the online experiments only the precision is used for measuring the performance. In the online experiment, rejection and verification modules are used. Results were averaged over 5 trials and every trial lasted 3 hours.

TABLE II: PRECISIONS OF ONLINE EXPERIMENTS

Approach	Precision (%)
EAR-SVM	82.0
EAR-GMM	75.7
EAR-NN	79.1
EAR-MV	90.1
EAR-FC	93.2

Table II presents the results among the systems with individual and ensemble classifiers in the online experiments. The table shows there is an average 12% increase for the system with ensemble classifiers. The proposed approach achieves an online precision of about 93.2% which fulfils the practical requirement.

VII. CONCLUSIONS

In order to protect children from adult videos, this paper proposes a novel and effective approach that detect adult videos. This has been taking into consideration that the feature porno-sounds in adult videos are inherently different from natural sounds. A novel unit called voiced fragments is proposed for classification. Individual classifiers are fusion classifier are trained.

The results of off-line and online experiments demonstrated that system using ensemble classifiers performs better than individual classifiers. The system achieves 93.2% precision rate in online test using ensemble classifiers with fusion classifier.

ACKNOWLEDGMENT

This work was partly supported by the National Natural Science Foundation of China under grant No. 91120303 and No. 61071181.

REFERENCES

[1] P. Y. Lee, S. C. Hui, and A. C. M. Fong, "Neural networks for web content fltering," *IEEE Intelligent Systems*, 2002, vol. 17, no. 5, 2002, pp. 48-57.

[2] M. Hammami, Y. Chahir, and L. Chen, "WebGuard: a Web fltering engine combining textual, structural, and visual content-based analysis," *IEEE Transactions on Knowledge and Data Engineering*, vol.18, no. 2, 2006, pp. 272-284.

[3] Z. Q. Shi, B. Y. Gao, J. Q. Han, and Z. Wu, "A study of objectionable sound recognition based on histogram features and SVM," in *Proc. of the International Congress on Image and Signal Processing*, 2009, pp. 1-4.

[4] Z. Q. Shi, B.Y. Gao, T. R. Zheng, and J. Q. Han, "Objectionable audio content recognition based on in-Class Clustering method," in *Proc. of the Network Infrastructure and Digital Content*, 2009, pp.712-716

[5] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks*, vol.14, no.1, 2003, pp. 209-215.

[6] X. Zhuang, X. Zhou, and T. S. Huang, "Feature analysis and selection for acoustic event detection," in *Proc.ICASSP*, 2008, pp.17-20.

[7] J. J. Aucouturier and F. Pachet, "Representing musical genre: A state of the art," *Journal of New Music Research*, vol. 32, no. 1, pp. 83–93, 2003.

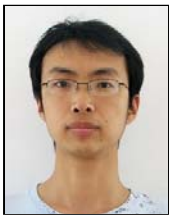
[8] T. Li and M. Ogihara, "Music genre classification with taxonomy," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*, March 2005, vol. 5, pp. 197–200.

[9] G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Transactions on Speech and Audio Processing*, vol. 10, no. 5, pp. 293–302, 2002.

[10] J. J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao, "A New Multiple Kernel Approach for Visual Concept Learning," in *Proc.MMM*, 2009.

[11] S. Sonnenburg, G. Raetsch, and C. Schaefer, "A general and effcient multiple kernel learning algorithm," *Advances in Neural Information Processing Systems*, 2005.

[12] I. H. Witten and E. Frank. "Data Mining: Practical Machine Learning Tools and Techniques," *Morgan Kaufmann, second edition*, 2005.



Ziqiang Shi received the B.S. degree in computer science from Northeastern University, China, in 2006 and the M.S. degree in computer science from Harbin Institute of Technology, China, in 2008. Presently he is pursuing a Ph.D. degree at Harbin Institute of Technology in computer science.

His research interests include audio information processing, speech signal processing, as well as pattern recognition and machine learning.

Tieran Zheng is an associate professor at School of Computer Science and Technology in Harbin Institute of Technology.

His research interests include speech signal processing and speech document retrieval.

Jiqing Han is an Associate Dean of School of Computer Science and Technology, Harbin Institute of Technology, Member of IEEE, Member of editorial board of Journal of Chinese Information Processing, Member of editorial board of Journal of Data Acquisition & Processing. Prof. Han is undertaking several projects from National Natural Science Foundation, 863Hi-tech Program, National Basic Research Program. He has won three Second Prize and two Third Prize awards of Science and Technology of Ministry/Province. He has published more than 100 papers and 2 books.

His research fields include speech signal processing and audio information processing.

Boyang Gao received the B.S. degree in computer science from Xi'an JiaoTong University, China, in 2006 and the M.S. degree in computer science from Harbin Institute of Technology, China, in 2009. Presently he is pursuing a Ph.D. degree at Ecole Centrale de Lyon in computer science.

His research interests include music information retrieval, speech signal processing, as well as machine learning.