Violence Detection in Movies with Auditory and Visual Cues

Jian Lin\textsuperscript{1} \quad Yi Sun\textsuperscript{1} \quad Weiqiang Wang\textsuperscript{1,2}

\textit{School of Information Science and Engineering, Graduate University of Chinese Academy of Sciences, Beijing, China}\textsuperscript{1}
\textit{Key Lab of Intell. Info. Process., Institute of Computing Technology, CAS, Beijing, China}\textsuperscript{2}

Email: \{jlin@jdl.ac.cn, ysun@gucas.ac.cn, wqwang@ict.ac.cn\}

Abstract—In this paper, we present a novel method to detect violent scenes in movies. The detection process involves two views—audio and video views. For the audio view, a supervised method based on HCRFs is exploited to improve the classification performance. For the video view, we detect violent shots by locating the violent areas. Finally, the auditory and visual classifiers are combined in a co-training way. The experimental results on several movies with violent contents preliminarily show the effectiveness of our method.

Keywords-HCRF; violent shots detection; audio-view; video-view; co-training

I. INTRODUCTION

The flourishing movie industry produced thousands of fascinating movies every year. However, not all the contents are suitable for children to watch, due to violent plots. Since more and more psychological researches 1 have shown the relationship between violent scenes and the realistic adolescent violent crimes, it is necessary to find an effective way to automatically detecting violent scenes in movies.

Some approaches have been proposed to address this problem. Swanson et al. 2 implemented the automatic conversion of movies with an ‘R’ rating to a ‘PG’ rating by hiding violent scenes using video data hiding techniques. Vasconcelos et al. 3 performed information filtering and genre classification in movie databases. However, more researches focus on detecting specific violent events. For instance, Datta et al. 4 exploited the accelerate motion vector to detect fist fighting, kicking. Nam et al. 5 located violent scenes by detecting flame and blood, and various audio effects, such as gunshots, explosions. Cheng et al. 6 proposed a hierarchical approach to recognizing gunshots, explosions, and car-braking. Generally speaking, movies do not have explicit structures like news and sports videos due to the aesthetic needs in movie-producing procedure. But there are still some common ‘film grammars’ for violent movies, and both auditory and visual cues can be utilized to detect the violent contents.

In our method, a video sequence is first split into a set of shots. Then the classification of these shots is implemented from two independent views— audio-view and video-view. For the audio-view, the audio segment of a specific shot is classified into violence or non-violence categories in a supervised way based on hidden conditional random field (HCRF). For the video-view, the corresponding video segment is classified by identifying the violent areas. Finally, the outputs of the two views are integrated to complement together through a co-training way. The proposed system is shown in Figure 1.

Figure 1. Framework of Our Approach

The rest of this paper is organized as follows. The proposed method is described from Section II to IV. In Section V, the experimental results are reported. Finally, Section VI concludes the paper.

II. AUDIO VIOLENCE DETECTION

In our system, an audio classifier is built in a supervised way based on HCRF, and is utilized to classify audio segments into violence or non-violence categories.

HCRF, as a model-based algorithm, was first applied in object recognition, and then extensively applied in various classification tasks, such as speech signal modeling, gesture recognition and ECG signal classification 7 8 9. The good performance has been shown in the above applications. In our system, the HCRF model is applied to locate the audio segments associated with violent contents. The class label $z$ for an input audio segment is predicted based on the feature sequence $a$ extracted from it, where $z \in Z = \{0,1\}$, corresponds to non-violence and violence categories respectively, and $a \triangleq \{a_j | j = 1,2,\ldots,m, a_j \in \mathbb{R}^d\}$, is a sequence of audio features for $m$ short-time windows.
Then the training set is $T = \{(a_i, z_i)| i = 1, 2, \ldots, n, z_i \in Z, a_i = \{a_{i,1}, a_{i,2}, \ldots, a_{i,d}\}\}$. For each observation variable $a_i$, there is a hidden audio-effect category $w \in W \triangleq \{w_1, w_2, \ldots, w_m\}$ labeling the corresponding audio segment, and many audio-effect categories in movies can be included, e.g., silence, music, speech, explosion, gunshot, yelling, car-braking. With observation $a$, label $z$, and hidden variable $w$, we define the audio HCRF model as

$$P(z|a, \theta) = \sum_w P(z, w|a, \theta) = \sum_w \frac{\exp(\psi(z, w, a; \theta))}{M(a; \theta)},$$

(1)

where $\theta$ are the parameters of the model, $M(a; \theta)$ is the normalized function defined by

$$M(a; \theta) = \sum_{w, z} \exp(\psi(z, w, a; \theta)), \quad (2)$$

and the potential function $\psi(z, w, a; \theta)$ with parameter $\theta \triangleq (\alpha, \beta)$ is defined as

$$\psi(z, w, a; \theta) = \sum_{k} \alpha_k f_k(z, w_1, a) + \sum_{k \geq 1} \beta_k g_k(z, w_{1-1}, w_1, a)], \quad (3)$$

where $f_k(z, w_1, a)$ and $g_k(z, w_{1-1}, w_1, a)$ represent the $k$th feature functions depending on one single hidden variable and two adjacent hidden variables respectively, and $\alpha_k, \beta_k$ are the corresponding parameters. The estimation of parameter $\theta$ of HCRF can be implemented through the objective function of CRF in [10], i.e.,

$$L(\theta) = \sum_{i} \log P(z_i|a_i, \theta) - \frac{1}{2\Delta^2} \| \theta \|^2. \quad (4)$$

The first term in (4) is the log likelihood function of data, and the second term is the log of a Gaussian prior with variance $\delta^2$. We use Quasi-Newton gradient ascent strategy to seek the optimal parameter $\theta^* = \arg \max_{\theta} L(\theta)$.

According to the data requirement of HCRF, we suppose a collection of audio documents is given. These segments are split into audio clips with 1 second long. In this work, the sampling rate is 44100 and 22050. We choose 512 samples as the length of short-time audio frame, and adjacent audio frames have 256 overlapping samples. Each clip is represented by a vector composed of local features, including spectrum power, brightness, bandwidth [11], pitch, MFCC, spectrum flux, high zero cross rate ratio and harmonic prominence [12], due to their success in speech recognition and audio classification [13].

In the training phase, given the sequence length and the number of hidden states, feature vectors and corresponding labels are fed into HCRF for modeling. In the testing phase, feature vectors extracted from testing sequence are fed into the model, then the category information is output to complement the video part through a co-training way, as the following computing procedure (Alg.1).

### Alg.1 Audio violence detection based on HCRF

**Step 1:** Substitute Eqn (1) into Eqn (4), we have

$$L(\theta) = \sum_{i} \log \frac{P(z_i, w_i, a_i; \theta)}{M(a_i; \theta)} - \frac{1}{2\Delta^2} \| \theta \|^2. \quad (5)$$

**Step 2:** Calculate the derivative of $L_i(\theta) = \log P(y_i|x_i; \theta)$ for a single training sample with respect to $\alpha_k$ and $\beta_k$,

$$\frac{\partial L_i(\theta)}{\partial \alpha_k} = \sum_{t \geq 1} \sum_{a_i} P(w_t|z_i, a_i; \theta) f_k(z_i, w_t, a_i) - \sum_{t, a_i'} P(w_t, z'|a_i) f_k(z_t', w_t, a_i) \quad \quad (6)$$

$$\frac{\partial L_i(\theta)}{\partial \beta_k} = \sum_{t \geq 1} \sum_{a_i} P(w_{t-1}, z_i, a_i; \theta) g_k(z_i, w_{t-1}, w_t, a_i) - \sum_{t, a_i'} P(w_{t-1}, w_t, z'|a_i) g_k(z_t', w_{t-1}, w_t, a_i) \quad \quad (7)$$

**Step 3:** Calculate (6) (7) by confidence delivery. With $\theta^*$ obtained by training data and given testing sequence $a$, we can get label $z$ of the audio segment $a$ by argmax$_{z \in Z} P(a|z, \theta^*)$

### III. VIDEO VIOLENCE DETECTION

According to film grammars, violent shots are always accompanied by violent events or violent objects. In this section, we locate violent shots by identifying violent objects.

Generally speaking, there are two ways to analyze violent objects. The first one is to isolate a single object, which always needs a priori knowledge and involves the image segmentation problem. The second way is to handle all violent objects as a whole—‘violent areas’, which can avoid the image segmentation problem and the requirement of a priori knowledge, and has a better generality.

In this work, we first sample the frames in the training set to perform violent block segmentation of violent areas. Then, for each violent block, we extract its local invariant features and perform clustering. Finally, each violent block is represented as a set of weighted clustering centers of local invariant features (feature signature). All the violent blocks obtained from the training set form a template library in 3 scales. In the testing phase, we identify violent areas by calculating the matching scores between the testing frames and templates through a modified Earth Mover Distance (EMD) [14], and then classify the frame as violence or non-violence. The video violence detection process is shown in Figure 2.

**A. Data Preparation**

We manually sample 1/5 of a movie to be tested and choose the frames which contain a large violent area or violent characters to perform ‘violent area’ segmentation. Then for each chosen frame, we find the violent content area and segment it into blocks by three scales: 32 by 32, 16 by 16, 8 by 8 pixels.

In our experiments, we find that it is more effective to enclose the violent areas accurately than to enclose them,
roughly. For example, for the violent character annotation, we can get better performance by enclosing big blocks (e.g., head, body) in the profile of the character than enclosing the character roughly. The reason may be this way reduces the noise of different background, so the performance of recognizing the same objects in different backgrounds is improved.

B. Violent Area Detection

Compared with the sports and news videos that have basically fixed scenes, illumination, camera motion patterns, etc., the movies have much more flexible and complex shooting angle and lighting scenes. So the features we choose to represent the violent areas should be robust to common geometric, photometric changes and partial occlusions, viewpoints and lighting changes.

In recent years, the effective, local invariant features are very popular in image and scene matching [15] [16] [17], and we exploit the method presented in [18] in this work to implement the matching between testing blocks and template blocks. For the testing and template blocks in the same scale, we use EMD to evaluate their matching score. The feature signatures \( B_p = \{p_i, w_{p,i} | i = 1, 2, \ldots, k \} \) and \( B_q = \{q_i, w_{q,i} | i = 1, 2, \ldots, l \} \) respectively represents the template block and the testing block [18], and the EMD between them is defined as

\[
EMD(B_p, B_q) = \min_{f_{ij}} \sum_{i,j} f_{ij} d(p_i, q_j)
\]

\[
s.t. \quad \sum_{i,j} f_{ij} = \min(\sum_i w_{p,i}, \sum_j w_{q,j}) \tag{8}
\]

\[
\sum_{j} f_{ij} \leq w_{p,i}, \quad \sum_{i} f_{ij} \leq w_{q,j}
\]

\[
f_{ij} \geq 0, \quad i = 1, 2, \ldots, k, j = 1, 2, \ldots, l
\]

where \( d(p_i, q_j) \) is the distance between feature vectors \( p_i \) and \( q_j \), and the distance is measured by

\[
d(p_i, q_j) = 1 - \sum_{p=1}^{D} \min(p_i^p, q_j^p) \tag{9}
\]

where \( p_i^p \) and \( q_j^p \) are respectively the \( p \)th entry of \( D \)-dimensional feature vectors \( p_i \) and \( q_j \). To boost discriminative power of EMD measure, we do non-linear scaling as [18], i.e.,

\[
d'(p_i, q_j) = \begin{cases} 0, & d(p_i, q_j) \in (0, \delta_1) \\ d(p_i, q_j), & d(p_i, q_j) \in (\delta_1, \delta_2) \\ d(p_i, q_j) \ast a, & d(p_i, q_j) \in (\delta_2, 1) \end{cases} \tag{10}
\]

where \( 0 < \delta_1 < \delta_2 < 1, a > 1 \).

According to the user attention theory and film grammar, only when the ratio between special contents and the whole frame exceeds a certain threshold, the users will pay great attention to the special contents. So we calculate all the matching scores of testing blocks in the testing frame, when the statistics exceeds a certain threshold, the frame is classified as violence. On the shot level, the evaluation value of the \( n \)th shot which contains \( n \) frames is defined as:

\[
E_m = \frac{\sum_{k=0}^{n-1} e_k}{n} \tag{11}
\]

where \( e_k \) is the violence score of the \( k \)th frame. If \( E_m \) exceeds a specified threshold \( \eta \), the \( m \)th shot is labeled as violence. The process of violent area detection is summarized in Alg.2.

## Alg.2 Violent Area Detection

**Repeat**

1. **Step 1:** Scan each testing frame using 32*32,16*16,8*8 sliding windows to get its feature signature as [19]
2. **Step 2:** Calculate the modified EMD measure between the feature signatures of the testing block and the blocks of the same scale in the violent template library.
3. **Step 3:** Calculate all the matching results of the testing blocks, if the ratio between the matched blocks and the testing blocks exceeds a certain threshold, the frame is classified as violence.
4. **Step 4:** Calculate matching results of all the testing frames in one shot, if the score exceeds a certain threshold, the shot is classified as violence.

**Until** all the data are labeled

IV. CO-TRAINING

Co-training algorithm works by generating several classifiers which are trained on a few labeled data. Then these classifiers are used to tag new unlabeled data. From the newly labeled data, the most confident predictions are sought, and subsequently added to the set of labeled data.
The process may continue for several iterations. One important aspect of co-training consists in the relation between the views used in learning. In the original definition of co-training [19], conditional independence of the views was considered as a required criterion for co-training. In the further work, co-training was still effective under a weaker independence assumption. A greedy algorithm was proposed to maximize agreement on unlabelled data, which produced good results in a co-training experiment. Moreover, [20] showed that similar performance can be achieved from a naive co-training process that did not seek to maximize agreement on unlabelled data, and with a much lower computational cost. After that, co-training was applied in statistical parsing [21], reference resolution [22], etc.

In our work, co-training is utilized in violence detection. Two different views of audio $\chi_a$ and video $\chi_v$ are involved in learning. Two classifiers $(C_a, C_v)$ are respectively built by HCRF in the audio view and in the video view. Every time, the two classifiers select the most likely $p$ positive examples, and the most unlikely $n$ negative examples. The details of co-training are described in Alg. 3.

**Alg. 3 Co-training learning algorithm**

input: $L$, the set of labeled training examples
$U$, the set of unlabeled examples

Output: the classifier $f$

Step 1: Create a pool $U'$ of examples by choosing $P$ random examples from $U$

step 2: while ( the unlabeled data is not used up)

(i) Use $L$ to individually train the classifier that considers only the audio feature view of the movie.

(ii) Use $L$ to individually train the classifier that considers only the video feature view of the movie.

(iii) For each classifier, select $p$ most confidently positive examples and $n$ most confidently negative examples from $U'$.

(iv) Add these $(p + n)$ examples to $L$.

(v) Refill $U'$ with examples from $U$, to keep $U'$ always has $P$ examples.

In our work, $L$ is formally defined as a data set which contains a collection of labeled shots randomly selected from several movies. The threshold value $\eta$ in video violence detection is firstly obtained from the labeled violent shots in $L$, and then is automatically adjusted by the labeled violent shots during the co-training process. For the $m$th shot, the threshold $\eta$ is calculated by the 10 latest labeled violent shots, i.e.,

$$\eta = 1.2 \times \min_{t=1}^{10} E_{m-t}$$

(12)

where $E_{m-t}$ is the video violence evaluation value of the $(m - t)$th shot.

V. EXPERIMENTAL RESULTS

The performance of the proposed approach is tested on three movies, which almost cover all kinds of dominant violence presentation styles in Hollywood movies. The detailed information is listed in Table II.

**Table II**

**EXPERIMENTAL DATASET INFORMATION**

<table>
<thead>
<tr>
<th>ID</th>
<th>Movie Title</th>
<th>Duration</th>
<th>Violence Type</th>
<th>Test Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Terminator I</td>
<td>108min</td>
<td>Gun/Explosion</td>
<td>53min.</td>
</tr>
<tr>
<td>2</td>
<td>Kill Bill I</td>
<td>111min</td>
<td>Fighting</td>
<td>36min.</td>
</tr>
<tr>
<td>3</td>
<td>Rock I</td>
<td>136min</td>
<td>Gun/Explosion/Fighting</td>
<td>55min.</td>
</tr>
</tbody>
</table>

Three measurements, i.e., precision $P$, recall $R$ and $F_1$ are utilized to evaluate our system, and they are defined by $P = \frac{n_T}{n_p}, R = \frac{n_T}{n_t}, F_1 = \frac{2P \times R}{P + R}$, where $n_T$ denotes the number of correctly detected violent shots, $n_p$ represents the number of shots declared as violence, and $n_t$ is the number of shots labeled as violence manually. $F_1$-measure is a harmonic mean of precision and recall, and is used to measure the overall performance of the method.

To verify the validity of the proposed approach, the comparison between SVM and our method is summarized in Table III. All three measurements of our method are much better than those generated by SVM. One possible reason should be the SVM treat visual and auditory features in the same way, and neglect the different classification effects of different views.

**Table III**

**RESULTS OF OUR METHOD AND SVM**

<table>
<thead>
<tr>
<th>ID</th>
<th>SVM</th>
<th>Our Method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>1</td>
<td>0.604</td>
<td>0.726</td>
</tr>
<tr>
<td>2</td>
<td>0.612</td>
<td>0.783</td>
</tr>
<tr>
<td>3</td>
<td>0.594</td>
<td>0.709</td>
</tr>
</tbody>
</table>

Even so, the results are still not good enough. One reason may be that even though we predict the HCRF of audio-view can get good performance for its time-dimension information, the empirical parameter-setting (the number of hidden states, the length of testing sequence, etc.) still affect the classification performance a lot. Other reasons may be that the annotation and scaling rules are still not the optimal solutions, and the representative features and the blocks search pattern still have important influence on the performance.

VI. CONCLUSIONS

In this paper, a new violent shot detection scheme is presented. From the audio-view, the violent shot is detected with the HCRF model. From the video view, the violent shot is detected by locating violent areas. Finally, the proposed audio and video classifiers are combined by the co-training.
technique. Experimental results show that the proposed method is effective in violent shots detection. Furthermore, the computational complexity is greatly reduced since a weakly-supervised method is adopted. In the future, more effective audio-video cooperation will be further investigated. With seeking for more effective feature extraction, representation, annotation and scaling rules, as well as block search pattern, we expect to obtain better performance.

ACKNOWLEDGMENT

This work was supported in part by National Key Technologies R&D Program under Grant 2006BAH02A24-2 and by National Natural Science Foundation of China under Grant 60873087.

REFERENCES


