Based on Multi-Modal Violent Movies Detection in Video Sharing Sites

Xingyu Zou\textsuperscript{1}, Ou Wu\textsuperscript{2}, Qishen Wang\textsuperscript{2}, Weiming Hu\textsuperscript{2}, Jinfeng Yang\textsuperscript{1}

\textsuperscript{1}College of aviation automation, Civil Aviation University of China, Tianjin, China
\{xyzou_yjs10, jfyang\}@cauc.edu.cn
\textsuperscript{2}National laboratory of pattern recognition, Chinese Academy of Sciences, Beijing, China
\{wuou, qswang, wmhu\}@nlpr.ia.ac.cn

Abstract. In this paper we present a method for the detection of violent movies in video sharing sites. The proposed method operates on three modalities: text, video and audio, the former being collected from the accompanying synopsis and user comments. Towards our goal, a multi-step approach is followed: initially, the text information is utilized to build a pre-classifier which selects the potential violent movie segments. At a second stage, a classifier is adopted, which combines the visual and audio information, in order to classify the potential violent movie segments as “violent” or “non-violent”. The experimental results on 220 movie segments from YouKu and TuDou show the effectiveness of our method.

Keywords: Violent movies Detection, Multi-modal, Text, Audio, Video.

1 Introduction

Movies are an important part of the daily entertainment. With the rapid development of the Internet technology, people increasingly prefer to watch movies in video sharing sites instead of the cinema. A huge amount of movies can be easily uploaded and accessed by online users of all ages including children, teenagers, etc. However, violence in movies has harmful influence on children. So it is necessary to protect children from accessing violent content. But the manually labeling huge volumes of video data is a hard work.

In this paper, we present an automatic method to detect violent movies in video sharing sites by combining text, video and audio information. Our method is divided into two stages. At the first stage, a text classifier is built as a pre-classifier, using the synopsis and the user comments on each movie online. The pre-classifier is used to select potential violent movie segments. Furthermore, the potential violent movie segments are segmented to a set of shots, using a shot detection method. Then four video features and two audio features are extracted from each shot. The final binary problem (i.e., violent vs non-violent movies) on a 6-dimensional feature space is solved by using Support Vector Machine (SVM).
The rest of the paper is organized as follows. Section 2 provides a brief overview of the related work. In Section 3, we detail the proposed method. Experiment results and discussions are presented in Section 4. Section 5 concludes this paper and gives the future work.

2 The related work

A few approaches have been proposed to violence detection in videos, owing to ambiguity in the definition of violence. It is difficult to give a precise definition about violence. Each related work has its own definition of violence. Current techniques for violence detection can be basically classified into three categories. The first one is based on visual cues. Using an accelerate motion vector, Datta et al. [1] addressed the problem of detecting human violence in videos such as fist fighting and kicking. The second category is based on audio cues. Cheng et al. [2] presented a hierarchical approach to locate gunplay and car racing. Giannakopoulos et al. [3] used eight audio features, both from the time and frequency domain, as input to a binary classifier which decides the video content with respect to violence. They proposed a multi-class classification algorithm for audio segments from movies in [4]. The third category is based on the fusion of visual and audio features. Nam et al. [5] combined multiple audio-visual features to identify violent scenes, in which flames and blood were detected by matching predefined color tables, and various representative audio effects (gunshots, explosions, etc.) were also exploited. Smeaton et al. [6] combined audio-visual features to select representative shots in an action movie to produce a trailer. Giannakopoulos et al. [7] presented a method for violence detection in movies based on audio-visual information that used a statistics of audio features and average motion and motion orientation variance features in video combined in a k-Nearest Neighbor classifier to decide whether the given sequence is violent. In this work, we research three typical events generally associated with violence, such as explosions, gunshots and fighting.

In this work, we use three modalities (text, video and audio) to detect violent movies. The use of text information can quickly filter out a large number of non-violent movies, reducing the cost of time. Moreover, the fusion of three modalities improves the detection accuracy compared with existing methods.

3 Our Method

Our method is composed of two stages, as shown in Fig. 1. In the first stage, a pre-classifier based text information is built to select the potential violent movie segments. In the second stage, we segment the video sequence into a set of shots. For each shot, video and audio features are extracted. Then a SVM classifier is trained to identify violent shots and movie segments are classified. The following subsections detail the two stages.
3.1 Detection of Potential Violent Movies

To a movie, the video sharing sites generally provide its synopsis and the users also give their own comments. In this paper, we consider the synopsis and comments as the text information associated with the movie. Compared with other types of movies, the movie types related to violence are usually action, war, crime and adventure, etc. We collect a dictionary of 250 keywords, which contains words frequently related to murder, crime, drug abuse, violence, war and weapon, etc. We have chosen Document Frequency (DF) in many text features, because it has low computational complexity and high computing speed. The processing steps of the text information are as follows.

The original text information is first preprocessed. Chinese word segmentation is carried out by ICTCLAS from ICT of Chinese Academy of Sciences [8]. Then the stop words are removed, such as preposition, conjunction and director, protagonist,
etc, which frequently occur in the synopsis and comments and are not useful for the classification.

In order to computing convenience, we adopt Vector Space Model (VSM) [9] to represent the text information because of its good performance. Each text $D_i$ is represented by a feature vector $(T_{1}, \omega_{1}; T_{2}, \omega_{2}; \cdots; T_{n}, \omega_{n})$ based on the dictionary of 250 keywords, where $T_{i}$ is the key word, $\omega_{i}$ reflects the degree of correlation between keywords and texts. The weight $\omega_{i}$ is computed by TF-IDF (Term Frequency-Inverse Document Frequency), it is defined as

$$
\omega_{i}(D,T_{i}) = \frac{1}{Z(T_{i},D)} \times T f_{i} \times i D f_{i} = \frac{1}{\sum_{j=1}^{M} (T f_{j})^{2} \times \log 2 \left( \frac{N}{N_{i}} + 0.01 \right)} \times T f_{i} \times \log \left( \frac{N}{N_{i}} + 0.01 \right)
$$

where $T f_{i}$ is the term frequency of the keyword $T_{i}$, $i D f_{i}$ is the inverse document frequency of the keyword $T_{i}$, $N$ is the total number of texts, $N_{i}$ is the number of the texts including the keyword $T_{i}$ and $M$ is the total of keywords. After VSM is built, the text vectors are normalized. Finally, a SVM classifier is trained to classify the movie segments as non-violent or potential violent ones. The process of pre-classification is presented in Fig. 2.

![Fig. 2. The process of pre-classification](image)

### 3.2 Video-Audio Classification

For each potential violent movie segment, we further check whether or not it is violent. First each video sequence is segmented into shots by the twin-comparison approach [10]. The shots whose duration is less than a threshold (e.g., 10 frames) are merged with their neighbor shots to avoid meaningless shots. Then each shot is
identified by the characteristics of violent movies. Film-making is a creative process, some universal rules should be followed. Generally, violent movies have the atmosphere of fast-pace and it is created by high-speed visual movement and fast-paced sound. Thus, we identify violent shot through the detection of fast camera movement, frequent shot transitions, sudden sound and typical events generally associated with violence, such as explosions, gunshots and fighting.

**Video Features.** The visual signal is split into shots. Each shot includes several video frames, and the visual features are calculated on every frame. Based on the visual characteristics of violent movies, the following features are extracted.

1) Motion Intensity:

Motion is an important visual feature which could describe the sustaining temporal variation of video streams. Also, it reveals the correlations between frame sequences within a video shot. Generally violent movie shots contain high activity and abrupt motion. To characterize the degree of motion within a shot, the average motion intensity is computed based on the motion vectors. First we compute the motion intensity of each frame. The $k$th frame of a shot is split into blocks of which we calculate the motion vectors using the previous frame as reference. The motion intensity of the $i$th block is defined as

$$M_k(i) = \sqrt{u^2(i) + v^2(i)}$$ (2)

where $(u_i, v_i)$ is the motion vector of the $i$th block of the $k$th frame. The average motion intensity of the $k$th frame is defined as

$$\overline{M_k} = \frac{1}{n} \sum_{i=0}^{n-1} M_k(i)$$ (3)

where $n$ is the number of blocks in the $k$th frame. Assuming there are $m$ frames in the shot, the average motion intensity of the shot is defined as

$$\overline{M} = \frac{1}{m} \sum_{k=0}^{m-1} \overline{M_k}$$ (4)

Finally, this value is normalized to be in the interval $[0, 1]$.

2) Flame:

Gunshot and explosion are typical and distinct violent events. One obvious visual feature is flame which suddenly generates in gunshot and explosion events. The flames have dominant yellow, orange and red color components. Thus, a predefined color template which includes a range of colors is employed to match flame colors. The flame of violent events is from scratch, and rapidly changes in a short period of time. This characteristic reflected on the image is that flame pixels rapidly increase or reduce within a few continuous frames. We are interested in the variation of frames with time, which reveal an obvious change in the number of flame pixels, rather than in the number of flame pixels owing to the existence of possible flame-like colors. The variation of flame pixels in a shot is defined as
where $F_i$ is the number of flame pixels in the $i$th frame, $M_f$ is the number of continuous frames whose flame pixels continue to change, and the value of $F_0$ is zero representing the neighbor frame without flame pixels before the flame frame. Finally the value $\bar{V}_f$ is normalized to be in the interval $[0, 1]$.

3) Bleeding:

Three typical violent events (fighting, explosion, gunshot) often lead to bleeding. Thus, bleeding could be a useful visual feature associated with violence. A predefined color template is employed to match bloody colors and bloody color pixels in video frames are identified. Similar to flame, the bloody element of violent events is from scratch, and rapidly increases in a short time. This characteristic reflected on the image is that bloody pixels rapidly increase within a few consequent frames. We focus on the variation of frames with time, which reveals an obvious increase in the number of bloody pixels, rather than in the number of bloody pixels owing to the existence of possible blood-like colors. The variation of bloody pixels in a shot is defined as

$$\bar{V}_b = \frac{1}{M_b} \sum_{i=0}^{M_b-1} |B_{i+1} - B_i|$$

where $B_i$ is the number of bloody pixels in the $i$th frame, $M_b$ is the number of sustained frames whose bloody pixels continue to increase, and the value of $B_0$ is zero representing the neighbor frame without bloody pixels before the bloody frame. Finally the value $\bar{V}_b$ is normalized to be in the interval $[0, 1]$.

4) Shot Length:

In order to attract attention, violent movies generally create a tense atmosphere. One of the most common methods is frequent conversion of the lens. As a result, the length of shots containing violence is commonly shorter than normal ones. That means violent shots have less frames than normal ones. Thus the length of shots also becomes a feature related to violence. The number of frames in a shot $L$ represents the length of the shot. Likewise, the value of $L$ is normalized to be in the interval $[0, 1]$.

Audio features. To exhibit exciting scenes we create the atmosphere of fast-tempo through not only visual sense but also auditory sense in violent movies. Generally, their audio characteristics include less speech, fast-paced music and typical audio events associated with explosions, gunshots and fighting. The common characteristics are the sound of intense fast-paced and short-term severe variations. Based on the visual characteristics of violent movies, the following features are extracted.

5) Audio Energy:

In violent shots, the sounds of fast-paced music, fierce fighting, explosion and gunshot provide additional energy to the shots. Compared with normal movies, the sounds in violent shots have often higher energy. Thus, the audio energy also
becomes a useful feature related to violence. The energy of the $i$th audio frame is defined as

$$ E(i) = \sum_{n=1}^{N} x_i^2(n) $$

where $N$ is the number of sampling points in the $i$th frame, $x_i(n)$ is the value of the $n$th sampling point in the $i$th frame. Assuming there are $m$ audio frames in the shot, then the average audio energy of the shot is defined as

$$ \bar{E} = \frac{1}{m} \sum_{i=1}^{m} E(i) $$

Finally, the value of $\bar{E}$ is normalized to be in the interval $[0, 1]$.

6) Energy Entropy:

In violent shots there are unique sound effects of beating, gunshot, explosion, crashing of objects and etc. They are mostly accompanied with a sudden burst in the audio level. We consider the abrupt change in the audio signal energy as another feature associated with violence. The energy entropy is employed to describe the abrupt change and defined as [11]

$$ I_n = - \sum_{i=1}^{J} \sigma_i^2 \log \sigma_i^2 $$

where $J$ is the total number of segments in the $n$th audio frame (each frame is divided into smaller $J$ segments) and $\sigma_i^2$ is the normalized energy of the $i$th short segment in a frame, $J$ is chosen to be 10. The value of the energy entropy measure falls down in frames with sudden energy transitions while the energy entropy is largest in a frame with nearly constant energy. Then, the minimum of the energy entropy in a shot is defined as

$$ I = \min_{n=1,\ldots,k} I_n $$

where $k$ is the number of audio frames in a shot. Finally, $I$ is normalized to be in the interval $[0, 1]$.

**Binary Classification.** The SVM classification model is adopted to solve this binary classification problem (violent vs non-violent content), where $S_i$ is the feature vector of the $i$th video and $R_i$ is the corresponding label. $N$ shot objects make up of the SVM training data, i.e., $N$ pairs of the form $(S_i, R_i)$, $i = 1, \ldots, N$. After all shots of a movie segment are detected, the movie segment is classified by a predefined threshold $T$, experimental results reach the best performance when the threshold is 0.1, i.e., when $P > T$, the movie segment is classified as “violent”, where $P$ is the percentage of violent shots in a movie segment. The false alarms will become less through the use of the threshold $T$. 
4 Experimental Results

4.1 Dataset

220 movie segments from YouKu and TuDou are used in our experiment, which have been manually labeled and segmented into labeled shots. The normal movie segments include a wide range of categories consisting of comedies, dramas, documentaries, romances and dance films, etc. The total video duration is 634 minutes, the average video duration is 2.88 minutes and 40% of the videos are less than 1 minute long. 121 movie segments are annotated as “violent”. The distribution of the film genres is presented in Fig. 3.

![Graph showing the percentage of movie genres](image)

Fig. 3. The percentage of movie genres

4.2 The result of Experiments

In our method, SVM with RBF kernel is used as the pre-classifier and video-audio classifier. In the SVM training process, 5-fold cross-validation is executed. We measure the performance of the proposed method by precision $P$, recall $R$ and $F_1$-measure $F_1$.

$$
P = \frac{N_d}{N_d + N_f}, \quad R = \frac{N_d}{N_d + N_m}, \quad F_1 = \frac{2 \cdot P \cdot R}{P + R} \quad (11)$$

where $N_d$ is the number of correctly detected violent movie segments, $N_f$ is the number of false detection, $N_m$ is the number of missed detection. $F_1$ is a harmonic mean of precision and recall, which is used to measure the overall performance of the method.

Table 1 lists the experimental results of the pre-classifier. In the process of pre-classification a large number of non-violent movie segments are filtered out. Because of the speed of text processing, the cost of time is greatly reduced. Meanwhile, most of violent movie segments are retained to enter the second step.
Table 1. The performance of the pre-classifier

<table>
<thead>
<tr>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>88.89%</td>
<td>75.47%</td>
<td>81.63%</td>
</tr>
</tbody>
</table>

When only the video and audio information are used, shots are split from movie segments, and then they are detected by video-audio classifier as non-violent or violent. Table 2 shows the experimental results of shots classification.

Table 2. The performance of shots classification

<table>
<thead>
<tr>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>79.74%</td>
<td>78.71%</td>
<td>79.22%</td>
</tr>
</tbody>
</table>

After the threshold $T=0.1$ is used, the result of movie segments classification is improved. Table 3 shows the result of movie segments classification without the pre-classifier.

Table 3. The result of movie segments classification

<table>
<thead>
<tr>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>79.74%</td>
<td>80.64%</td>
<td>80.19%</td>
</tr>
</tbody>
</table>

Finally, all the three modalities are used. The result is shown in Table 4. Compared with the above results, the use of three modalities results in the precision improvement of 2.77%.

Table 4. The final result of our method

<table>
<thead>
<tr>
<th>Recall</th>
<th>Precision</th>
<th>$F_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>84.22%</td>
<td>83.41%</td>
<td>83.81%</td>
</tr>
</tbody>
</table>

5 Conclusion

In this work, we present an effective approach for automatically detecting violent movies in video sharing sites. While most of previous approaches addressed on one or two modalities only, our approach works on three modalities: text, video, and audio. Hence, more useful information of movies is utilized for detection. Text information is used to design a pre-classifier to identify potential violent movie segments and they are classified by video-audio classifier as violent or non-violent. Though the extracted features are simple, experimental results show that the proposed method is efficient and effective. In the future, we plan to combine other useful information (e.g., other
video information, other audio information and so on) to further improve the performance.

Acknowledgment

This work is partly supported by the National Nature Science Foundation of China (No.61005030, 60935002, and 60825204) and Chinese National Programs for High Technology Research and Development (863 Program) (No. 2012AA012503 and No. 2012AA012504).

References