HORROR VIDEO SCENE RECOGNITION VIA MULTIPLE-INSTANCE LEARNING

Jianchao Wang, Bing Li, Weiming Hu, Ou Wu

NLPR, CASIA, 95 Zhongguancun East Road, 100190, BEIJING, CHINA
{ericustc, bjtulb}@gmail.com, {wmhu, wuou}@nlpr.ia.ac.cn

ABSTRACT
Along with the ever-growing Web comes the proliferation of objectionable content, such as pornography, violence, horror information, etc. Horror videos, whose threat to children's health is no less than pornographic video, are sometimes neglected by existing Web filtering tools. Consequently, an effective horror video filtering tool is necessary for preventing children from accessing these harmful horror videos. In this paper, by introducing color emotion and color harmony theories, we propose a horror video scenes recognition algorithm. Firstly, the video scenes are decomposed into a set of shots. Then we extract the visual features, audio features and emotional features of each shot, the video scene is viewed as a bag and each shot is treated as an instance of the corresponding bag. Finally, by combining the three features, the horror video scenes are recognized by the Multiple-Instance learning (MIL). According to the experimental results on diverse video scenes, the proposed scheme based on the emotional perception could effectively deal with the horror video scene recognition and promising results are achieved.

Index Terms— Horror Movie Recognition, Affective Understanding, Color Emotion, Color Harmony, Multiple-Instance learning

1. INTRODUCTION
As a large, widely distributed, global information center, the World Wide Web is growing ever more rabidly. More and richer information sources and services are available on the Web everyday. However, along with the ever-growing Web comes the proliferation of objectionable content, such as pornography, violence, horror information, etc., which is not appropriate for all users, notably children. The openness of the Web allows any user to access almost any type of information, so we need efficient tools for classifying and filtering undesirable web content.

Recently, a variety of horror materials that are accessed by children easily are becoming a big issue. A number of psychological and physiological researches indicate that too much horror images or films can seriously affect childrens health. Rachmans research [1] shows that horror information is one of the most important factors for phobias. Field et al. [2] further point out that horror information can increase behavioral avoidance as well as fear beliefs. The experiments of King et al. [3] further indicate that 88.8% children ascribe their phobias to horror information acquisition.

Considering the harm of horror information, an effective horror information filtering tool is necessary. In the past decades, many studies have mainly focused on Web content filtering such as the pornographic content filtering [4] [5]. However, as far as we know, there is nearly no research on the horror content filtering. In this paper, by introducing color emotion and color harmony theories, we propose a horror video scene recognition algorithm which is the key ingredient of horror content filtering.

Horror movies, which are major component of horror messages, are films that strive to elicit the emotions of fear, horror and terror from viewers. How to bridge the gap between low-level visual-audio features and high-level affective understanding is a challenge for horror video recognition. Although there is nearly no special research on horror video recognition, research on affective content analysis in videos and audio is an established research area.

Affective scene classification, which categorizes movie scenes by the emotion, is an attractive framework for this issue. Most existing methods focus on detecting movie affective content by using low-level features [6,7,8,9,10]. Hee et al. [6] extract a number of effective audiovisual cues to help bridge the affective gap and introduce a holistic method of extracting affective information from the multifeatured audio stream. In [7], a computational framework for affective video content representation and modeling is proposed and the affective video content is mapped onto the 2-D emotion space. The emotion space is characterized by the dimensions of arousal (intensity of affect) and valence (type of affect), by using the models that link the arousal and valence dimensions to low-level features extracted from the video data. Zeeshan et al. [8] present a framework for the classification of films into genres, based only on computable visual cues. Kang use HMMs to categorize movie scenes into three types of affective content, joy, fear, and sadness, based on low-level visual features [9]. Xu et al. [10] proposed an HMM-based method to detect affective events such as laughter in comedies and terrible sunds in horror movies.

The remainder of this paper is organized as follows. Section 2 discusses the proposed scheme for horror film scene recognition, including the emotional feature based on Color emotion and color harmony theory. Section 3 shows our experimental results for horror movie scene recognition. Section 4 gives the conclusion and future work.

2. PROPOSED METHOD
The proposed approach consists of three main steps as illustrated in Figure 1. First, movie scene is divided into a series of shots via shot detection. Then audio stream and video key frame of each shot are characterized by audio features, the visual features and color emotional features respectively. Each shot is viewed as an instance of the movie scene, which is treated as a bag. Four feature combinations are obtained from the three types of different features. Finally, based on the extracted features, the horror movie scene is recognized by the multiple-instance learning method. The details of the components of the proposed approach are discussed in following the subsection.

Video segmentation is a fundamental step in analyzing video content. Cernekova et al. [11] used mutual information (MI) to measure information transported from one frame to another. Abrupt
transitions and fades between two shots lead to a low level of MI. This approach achieves an impressive performance on shot change detection, so we adopt this method to segment the movie scene into shots, and then the central frame of each shot is chosen as the key-frame in order to reduce computational complexity. Besides video segmentation, feature extraction is the most important step for horror video recognition. In order to represent the horror movie scene, visual features (VF), audio features (AF) and color emotional features (EF) are extracted.

2.1. Visual features

In practice, movie directors use multiple light sources to enable a specific portrayal of a scene. Lighting can also be used to direct the viewer to certain area of importance in the scene and can also affect the viewer’s feeling directly regardless of the actual content of the scene. Generally two major lighting techniques, low-key lighting and high-key lighting, are frequently employed. In the cinematographic perception, the sadness, fear, and surprise for sad, frightening, or suspense scenes are recreated by the use of dim lights, shadow play, and predominantly dark background [6][8]. Lighting key is determined by two factors: 1) the general level of light and 2) the proportion of shadow area. In order to detect the lighting key, two visual features are formatted in [8]. The first component can be characterized by using the median of the value \(L\) of the \(Luv\) color space of the key frame. The proportion of pixels, whose lightness are below a certain shadow threshold \(T\), is used as an indicator of the second component which is the proportion of shadow area. \(T\) is experimentally determined to be 0.18.

The HSV color space is quite similar to the way in which humans perceive and the colors used in HSV can be clearly defined by human perception, so we calculate the means and variances of the HSV color as the overall characteristics of the key frames. Intuitively, the variance of color has a strong relation with movie genres. For instance, comedies tend to have a large variety of bright colors, whereas horror films often adopt only darker hues [8]. To represent the variance of color used in the movie scene, we employ the generalized variance of color space of the key frames in the movie scene. The covariance matrix of \(L, u, v\) of each key frame is defined as:

\[
\rho = \begin{bmatrix}
\sigma_L^2 & \sigma_L^2 & \sigma_L^2 \\
\sigma_L^2 & \sigma_L^2 & \sigma_L^2 \\
\sigma_L^2 & \sigma_L^2 & \sigma_L^2
\end{bmatrix}
\]

(1)

The generalized variance is obtained by finding the determinant of Eq(1): \(\sum = \det(\rho)\). This feature is used as a representation of the color variance.

2.2. Audio features

It is well known that, in a sound film, movie editors usually use some specific sounds and music to highlight emotional atmosphere and promote dramatic effects. Although the emotional meaning of music is subjective and it depends on many factors including culture, it is also found that, within a given cultural context, there is an agreement among individuals as to the mood elicited by music. The mel-frequency cepstrum has proven to be highly effective in automatic speech recognition and in modeling the subjective pitch and frequency content of audio signals. The mel-cepstral features can be illustrated by Mel-Frequency Cepstral Coefficients (MFCCs), which are computed from the Fast Fourier Transform (FFT) power coefficients. For audio signal, we extract a single-channel audio stream at 44.1 KHz and compute 12 MFCCs over 20ms frames. The means and variances of both the 12 MFCCs of each frame and 12 MFCCs’ first-order differential are adopted in this paper.

For an audio signal \(s(n)\), each frame is weighted with a hammering window \(h(n)\), where \(N\) is the number of samples of each frame. The spectral power of the signal \(s(n)\) of the audio frame is calculated as

\[
\text{Power}(k) = 10\log \left[ \frac{1}{N} \left| \sum_{0}^{N-1} s(n)h(n)\exp(-j2\pi nk/N) \right|^2 \right]
\]

(3)

The spectral centroid of audio signal whose mean and variance are employed is a measure of spectral shape and higher centroid values correspond to “brighter” texture with more high frequencies. Time domain zero crossings rate provide a measure of the noisiness of the audio signal [13].

2.3. Color emotional features

Both visual features and audio features belong to low-level features that contain little high-level emotional perception. In this subsection, we extract some emotional features based color emotion and color harmony theories. Color emotion and color harmony are the...
high-level semantic concepts of images. Ou et al. [14] used psychophysical experiments to develop color emotion models for single colors and two-color combinations and investigated the relationship between color emotion and color preference. Color emotion model for single-colors are derived from psychophysical experiments, resulting in three color emotion factors: activity(A), weight(W) and heat(H):

\[ \begin{align*}
A &= -2.1 + 0.06 \left[ (L^* - 50)^2 + (a^* - 3)^2 + \left( \frac{b^* - 17}{1.4} \right)^2 \right]^{1/2} \\
W &= -1.8 + 0.04(100 - L^*) + 0.45\cos(h - 100^\circ) \\
H &= -0.5 + 0.02(C^*)^{0.07}\cos(h - 50^\circ)
\end{align*} \]

where \((L^*, a^*, b^*)\) and \((L^*, C^*, h)\) are the color values in CIELAB and CIELCH color spaces respectively. Given a key frame of video sequence, we convert its RGB color to CIELAB and CIELCH color spaces, followed by another computation according to Eq (4), its color emotion value is denoted as \(\text{ColorEmotion} = [\text{Activity}, \text{Weight}, \text{Heat}]\). We define single pixels emotional intensity \(EI\) as:

\[ EI(x, y) = \sqrt{\text{Activity}^2 + \text{Weight}^2 + \text{Heat}^2} \]  

According to Eq(5), we can get color emotion histogram of the key frame, which has 40 bins. The emotion histogram is employed as one part of the color emotional features.

In color research, a common definition of harmonious color combination combinations is “colors that are said to generate a pleasant effect when seen in neighboring area”. We utilize the quantitative two-color harmony model developed by Ou et al. [15] to derive color harmony scores. The model consists of three independent color harmony factors: chromatic effect\((CH_C)\), lightness effect\((CH_L)\), hue effect\((CH_H)\). The factors are combined to form a two-color harmony model, where \(CH\) defines the overall harmony score.

\[ CH = CH_C + CH_L + CH_H \]  

Since the equations for the three factors are very complex, we do not provide the entire equations here, the details can be found in [15]. We can obtain every pixel’s the color harmony score \(CH1\) between itself and its’ surrounding pixels and the color harmony score \(CH2\) between itself and the whole key frame’ pixels according to Eq(6). Every pixel’s color harmony score \(CH = 0.5(CH1 + CH2)\). Then we can get the color harmony histogram of the key frame. The color harmony histogram which has 40 bins is the other part of the emotional features.

2.4. Multiple-instance learning

Multiple-instance learning (MIL) is a generalization of supervised classification in which training class labels are associated with sets of patterns, or bags, instead of individual patterns. While every pattern may possess an associated true label, it is assumed that pattern labels are only indirectly accessible through labels attached to bags. The law of inheritance is such that a set receives a particular label, if at least one of the patterns in the set possesses the label. In the important case of binary classification, this implies that a bag is “positive” if at least one of its member patterns is a positive example. Horror video recognition is a multiple-instance problem. A video scene consist of a series of shots, and if at least one of the shots is a horror shot, the video scene is horror one. if all of the shots are non-horror, the video scene is non-horror one. Therefore we can treat the video scene as a bag and its’ shots as bag’s patterns. As described in previous sections, firstly, the video scene is divided into shots, then visual features, audio features and emotional features of each shot are extracted, last MIL is adopted to recognize the horror video scene. Besides, to compare with MIL, we convert the multiple-instance problem to single instance problem by obtaining the mean of the patterns’ features of the bag. The mean is treated as the feature of the bag, and then the Classifier Support Vector Machine (SVM) is employed to recognize the horror movie scene.

3. EXPERIMENTS

Fig. 2. Performance of Horror Video Recognition (A)Results of Horror Movie Scene Identification by MI-SVM (B)Results of Horror Movie Scene Identification by CKNN (C)Results of Horror Movie Scene Identification by EM-DD (D)Results of Horror Movie Scene Identification by SI-SVM

Fig. 3. \(F_1\) of different classifiers with different feature combinations

3.1. Data Set

We download a large number of movies from the internet. The film data collected from the Internet consist of 100 horror movies and 100 non-horror movies which are from different countries such as China, US, Japan, South Korea and Thailand etc. The genres of the
non-horror movies are comedy, action, drama and cartoon. We get 370 horror movie scenes and 370 non-horror movie scenes. The movie scenes are divided into subset A and subset B. A consists of 185 horror movie scenes and 185 non-horror movie scenes. B consists of 185 horror movie scenes and 185 non-horror movie scenes. In order to remove the correlation, we place the movie scenes which derive from the same movie in the same subset.

3.2. Results

Shots of the movie scene are represented by visual features (V.F,11D), audio features (AF,52D) and color emotional features (EF,80D). Seven different combinations of the features can be got from the three types of features. Then MI-SVM [16], citation K-Nearest Neighbor (CKNN) [17] and Expectation-Maximization version of Diverse Density (EM-DD) [18] are employed to recognize the horror movie scene in all cases of feature combination. Fig.2 (A),(B) and (C) show the results of horror movie scenes identification of MI-SVM, CKNN and EM-DD respectively. Common single-instance SVM (SI-SVM) also is employed to classify the horror movie scene as the comparative experiment and the results of the common SVM are shown in Fig.2(D). In experiments Precision (P), Recall (R) and F-measure (F1) are used to evaluate the recognition performances. The F-measure which is commonly used in information retrieval is defined as:

Experimental results in Fig.2 show that the best one among three types of features is the audio feature, which has the highest F-measure. We can find that regarding combinational features, combination of the visual features, audio features and color emotional features performs best, with F1 = 0.7535 in Fig.3, when combined with audio features, color emotional features with F1 = 0.7908 and visual features with F1 = 0.7862 play the same role in horror video scene recognition. In Fig.2, (A), (B) and (C) show that MI-SVM performs better than CKNN and EMDD in horror video scene recognition. We also find that MI-SVM is more suitable than single instance classifier SVM for horror movie scene recognition in Fig.3.

4. CONCLUSION

In this paper, we have proposed an effective approach to solve the problem of horror movie scene recognition, which is an initial stage for Web horror information filtering. Color emotion and color harmony theories are introduced into the horror video scene recognition and we extract color emotional features which are high-level features. And MIL is employed in horror movie scene recognition. Experimental results show that color emotional features can improve the horror movie scene recognition effectively. In future work, we will be dedicated to improving the horror film scene recognition algorithm by better modeling of features fusion.

Acknowledgments This work is partly supported by NSFC (Grant No. 60825024, 60935002, 60723005 and 61005030). The author Bing Li is also supported by China Postdoctoral Science Foundation and K.C.Wong Education Foundation.

5. REFERENCES