Context-aware Horror Video Scene Recognition via Cost-sensitive Sparse Coding

Xinmiao Ding
China University of Mining and Technology(Beijing)
dingxinmiao@126.com

Bing Li, Weiming Hu
National Laboratory of Pattern Recognition, Institute of Automation,CAS
{*bli,wmhu*}*@nlpr.ia.ac.cn*

Weihua Xiong
OmniVision Technologies,Sunnyvale,CA,USA

Zhenchong Wang
China University of Mining and Technology(Beijing)

Abstract

Along with the ever-growing Web, horror video sharing through the Internet has affected our children's psychological health. Most of current horror video filtering researches pay more attention to the extraction of global features or selection of an optimal classifier, while neglecting the underlying contexts in a scene. In this paper, a novel cost-sensitive sparse coding (CSC) model is proposed to address the context inside scene and interrelations between audio-visual features simultaneously. The model essentially includes two aspects: one is to construct inner contextual structure among frames from same scene based on a ϵ-graph; the other one is to extend the classic sparse coding technique into a cost-sensitive sparse coding model for graph pattern classification as well as audio-visual features fusion through graph kernel. The experiments on various video scenes demonstrate that our method's performance is superior to the other existing algorithm.

1. Introduction

With the rapid development of the Internet, the openness of the Web allows users to access almost all types of information, including pornography, violence, horror information,etc. These objectionable contects are not appropriate for all users, especially children. To protect our psychological health, lots of scientific researchers have investigated web filters to block objectionable contents automatically. Some of them, for example pornographic content filters, have matured to a point where robust recognition or filter software is available [2]. By contrast, the research of affective semantics of horror video is still on the stage of exploration. Therefore, an effective horror video scene recognition algorithm is necessary for web filtering.

1.1. Related work

The earlier work on horror video scene recognition can be dated back to a part of affective video scene classification whose final goal is to categorize movie scenes based on human emotions. The horror video is picked up if a type of fear emotion is recognized. Most existing video scene affective classification methods [10, 1, 8] focus on mapping low-level features to high-level emotions. Kang [4] introduces the Hidden Markov Model (HMM) to categorize movie scenes into three types of affective content: joy, fear, and sadness, based on low-level visual features. The latent topic driving model is also applied to affective scene classification [3]. As an emerging problem, horror scene recognition has its own characteristics. Several researchers begin to pay special attention to this area [11, 12, 14, 7]. Wang et al. [11] firstly use Support Vector Machines (SVM) to identify horror scene on several holistic effective features inspired by emotional perception theory. But they further find that the horror scene sometimes contain several rather than most horror frames. The holistic features inevitably weaken the features of the real horror frames. In order to avoid this confusion, the multi-instance learning (MIL) is also introduced by both Wang et al. [12] and Wu et al. [14], in which the scene is represented as a bag of key frames with corresponding independent features.

1.2. Our work

Either the holistic methods or MIL based methods only focus on independent frames, without taking in-

to account the underlying contextual cues in the video scene. However, as Li et al. [7] point, the horror emotion recognition should benefit from the proper use of contextual cues. The contextual cues in video scene mainly include two parts: context among frames belonging to the same scene and context between visual and audio cues. In order to effectively take advantage of the contextual cues, we propose a novel cost-sensitive sparse coding (CSC) model based on the graph kernel to represent these two contextual relationships for improving horror scene recognition. The framework of the proposed method is shown in Figue 1. First, a movie scene is divided into a series of shots via shot segmentation and the key frame of each shot is picked out. The visual feature of every key frame and the audio feature of the scene, rather than shot, are extracted. Now a scene can be represented as a bag of key frames with corresponding visual feature vectors and an audio feature vector. Second, the ϵ-graph is constructed among key frames to represent their contextual relationship. Finally, we extend the sparse coding technique by proposing a novel cost-sensitive sparse coding (CSC) model to represent the context between visual and audio features. Experimental results on various videos show that the proposed CSC method is superior to the state-of-the-art methods.

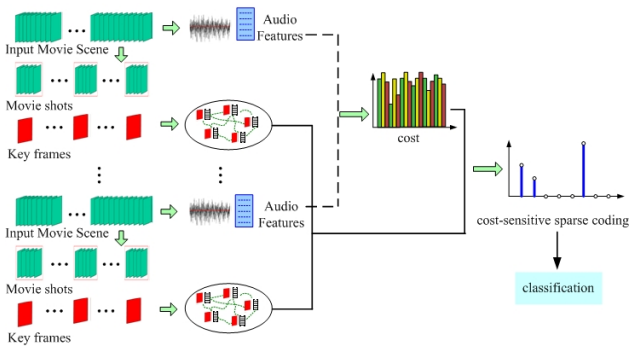

Figure 1. Framework of the proposed method

2. The proposed method

2.1. Video and audio feature representation

Given N training video scenes $V_1, V_2, \cdots, V_N$ with their labels $y_1, y_2, \cdots, y_N (y_i \in \{-1, +1\})$, each video scene V_i is divided into n_i shots $S_{i,1}, S_{i,2}, \cdots, S_{i,n_i}$, in which the key frames $F_{i,1}, F_{i,2}, \cdots, F_{i,n_i}$ are extracted. Then we extract the visual feature $v_{i,j} \in R^m$ of the frame $F_{i,j}$. In addition, the audio feature $a_i \in R^p$ of a scene V_i is also extracted. The audio feature of a scene rather than a shot is used in this paper due to the fact that a relatively long-time audio change expresses a certain emotion much better than a shorter one from a shot. MI(Mutual Information) based shot segmentation algorithms as well as visual and audio features in [12] are used in this paper.

2.2. ϵ-graph construction for visual context representation

Following the video structural representation is to define the relationships among key frames' visual features. Inspired by [15], the ϵ-graph, which is shown helpful [9] in discovering the underlying manifold structure of data, is used to model the context among key frames in each scene. For a scene V_i, $W^i \in R^{n_i \times n_i}$ is set as a ϵ-graph adjacency weight matrix. We compute the distance of every pair of key frame nodes, e.g. $v_{i,k}$ and $v_{i,l}$. If the distance between $v_{i,k}$ and $v_{i,l}$ is smaller than a pre-set threshold ϵ, then an edge is established between these two key frame nodes, and the weight value W^i_{kl} is set to 1, otherwise 0. According to the manifold property [9], i.e., a small local area is approximately an Euclidean space, we use the Euclidean distance $\|v_{i,k} - v_{i,l}\|$ between $v_{i,k}$ and $v_{i,l}$ to establish the ϵ-graph. Finally, a bag of visual feature vectors of V_i are reconstructed as a ϵ-graph G_i.

2.3. Cost-sensitive sparse coding for visual-audio context

Besides ϵ-graph representation among key frames, another important factor is from audio cues and its interrelation with visual features. To the end, we propose a cost-sensitive sparse representation on graph patterns. Given a test video scene V', its ϵ-graph is constructed as G' and the audio feature a' is extracted. Inspired by [6], we apply a feature mapping function φ: $G \rightarrow R^d$ to map the graph G to a higher dimensional feature space: $G \rightarrow \varphi(G)$. Thus, we can obtain a basis matrix $\mathbf{U} = [\varphi(G_1), \varphi(G_2), \cdots, \varphi(G_N)]$. The CSC coding is formulated in a high dimensional feature space as:

$$\min_{\beta} \|\varphi(G') - \mathbf{U}\beta\|^2 + \lambda\|\mathbf{D}\beta\|_1$$
$$\mathbf{D} = diag(\|a_1 - a'\|, \cdots, \|a_i - a'\|, \cdots, \|a_N - a'\|) \quad (1)$$

where the first term of Eq.(1) is the reconstruction error, and the second term is used to control the sparsity of the coefficient vector β with the l_1 norm. λ is regularization coefficient to control the sparsity of β. The larger λ implies the sparser solution of β. From Eq.(1),

we can find two differences between the CSC model and the general sparse coding model [13]: (1) the graph patterns are used in CSC, while general vectors are used in sparse coding. (2) A diagonal matrix $\mathbf{D}$ is added into the l_1 norm, which can be viewed as cost values to different training samples. Therefore, cost values are actually audio feature distances from the test scene to the training scenes. Minimization of Eq.(1) targets to select those training samples, which have lower audio feature distances from the test scene, to reconstruct the visual feature of the test scene. In other words, the CSC model considers both visual and audio cues simultaneously in the reconstruction procedure.

2.4. Optimization for CSC model

This section discusses how to optimize the objec-t function defined in Eq.(1). Let $\gamma = \mathbf{D}\beta$, then $\beta = \mathbf{D}^{-1}\gamma$, the Eq.(1) can be rewritten as:

$$\begin{aligned}&\min_{\gamma} \|\varphi(G') - \mathbf{U}\mathbf{D}^{-1}\gamma\|^2 + \lambda^{'}\|\gamma\|_1 \\ &\mathbf{D}^{-1} = diag(\frac{1}{\|a_1 - a'\|}, \cdots, \frac{1}{\|a_i - a'\|}, \cdots, \frac{1}{\|a_N - a'\|})\end{aligned} \tag{2}$$

If we set $\mathbf{V} = \mathbf{U}\mathbf{D}^{-1}$, Eq.(2) can also be rewritten as:

$$\min_{\gamma} \|\varphi(G') - \mathbf{V}\gamma\|^2 + \lambda^{'}\|\gamma\|_1 \tag{3}$$

,where

$$\|\varphi(G') - \mathbf{V}\gamma\|^2 = [\varphi(G')]^T\varphi(G') + \gamma^T\mathbf{V}^T\mathbf{V}\gamma - 2\gamma^T\mathbf{V}^T\varphi(G') \tag{4}$$

. The Eq.(3) is essentially a general sparse coding optimization problem. If $\mathbf{V}^T\mathbf{V}$ and $\mathbf{V}^T\varphi(G')$ are given out, the optimization in Eq.(3) can be easily and efficiently solved by recently proposed Feature-Sign Search algorithm (FSS) [5]. The Eq.(4) is actually equivalent to:

$$\begin{aligned}&[\varphi(G')]^T\varphi(G') + \gamma^T(\mathbf{D}^{-1})^T\mathbf{U}^T\mathbf{U}\mathbf{D}^{-1}\gamma - 2\gamma^T(\mathbf{D}^{-1})^T\mathbf{U}^T\varphi(G') \\ &= K_g(G', G') + \gamma^T(\mathbf{D}^{-1})^T \\ &\begin{bmatrix} K_g(G_1, G_1) & K_g(G_1, G_2) & \cdots & K_g(G_1, G_N) \\ K_g(G_2, G_1) & K_g(G_2, G_2) & \cdots & K_g(G_2, G_N) \\ \cdots & \cdots & \cdots & \cdots \\ K_g(G_N, G_1) & K_g(G_N, G_2) & & K_g(G_N, G_N) \end{bmatrix} \mathbf{D}^{-1}\gamma \\ &- 2\gamma^T(\mathbf{D}^{-1})^T \begin{bmatrix} K_g(G_1, G') \\ K_g(G_2, G') \\ \cdots \\ K_g(G_N, G') \end{bmatrix} \\ &= 1 + \gamma^T(\mathbf{D}^{-1})^T\mathbf{K}_{\mathbf{UU}}\mathbf{D}^{-1}\gamma^T - 2\gamma^T(\mathbf{D}^{-1})^T\mathbf{K}_{\mathbf{U}G'}\end{aligned} \tag{5}$$

where $K_g()$ is a graph kernel function that expresses the dot product of graphs in a high dimensional feature space. If both kernel matrixes $\mathbf{K}_{\mathbf{UU}}$ and $\mathbf{K}_{\mathbf{U}G'}$ are obtained, the $\mathbf{V}^T\mathbf{V}$ and $\mathbf{V}^T\varphi(G')$ can also be easily calculated. Consequently, the optimization of Eq.(3) can also be easily solved by FSS. Many existing graph kernel functions can be applied. We use the same graph kernel function in [15]:

$$\begin{aligned}&K_g(G_i, G_j) = \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}\omega_{i,a}\omega_{j,b}K(v_{i,a}, v_{j,b})}{\sum_{a=1}^{n_i}\omega_{i,a}\sum_{b=1}^{n_j}\omega_{j,b}} \\ &K(v_{i,a}, v_{j,b}) = exp(-\gamma\|v_{i,a} - v_{j,b}\|^2)\end{aligned} \tag{6}$$

where $\omega_{i,a} = 1/\sum_{u=1}^{n_i}\mathbf{W}^i_{a,u}$, $\omega_{j,b} = 1/\sum_{u=1}^{n_j}\mathbf{W}^j_{b,u}$, $\mathbf{W}^i$ and $\mathbf{W}^j$ are the adjacency weight matrixes for scene V_i and V_j, respectively. In addition, $K(v_{i,a}, v_{j,b})$ is defined using the Gaussian radial basis function (RBF) kernel.

2.5. Scene classification

After the coefficients vector γ is obtained, the reconstruction residual $r_q(G')$ of the test scene in class $q \in \{-1, 1\}$ is defined as:

$$\begin{aligned}&r_q(G') = \|\varphi(G') - \mathbf{U}\mathbf{D}^{-1}\delta_q(\gamma)\|^2 \\ &= 1 + \delta_q(\gamma)^T(\mathbf{D}^{-1})^T\mathbf{K}_{\mathbf{UU}}\mathbf{D}^{-1}\delta_q(\gamma) - 2\delta_q(\gamma)^T(\mathbf{D}^{-1})^T\mathbf{K}_{\mathbf{U}G'} \\ &[\delta_q(\gamma)]_k = \begin{cases} \gamma_k, & y_k = q \\ 0 & y_k \neq q \end{cases}\end{aligned} \tag{7}$$

where $\delta_q(\gamma)$ is a coefficient selector that only selects coefficients associated with class q. The final class c that is assigned to the test video scene V' is the one that gives the smallest residual, as:

$$c = arg\min_q(r_q(G')) \tag{8}$$

3. Experiments

3.1. Data set

We download from the internet a large number of movies which consist of 100 horror movies and 100 non-horror movies from different countries such as China, US, Japan, South Korea, and Thailand etc. The genres of the non-horror movies include comedy, action, drama and cartoon. We get 400 horror movie scenes and 400 non-horror movie scenes in total. The proposed method is compared with MIL-based horror video scene recognition methods proposed by Wang et al. [12]. In order to validate the effect of the audio cost, the CSC model without audio cost (denoted as SC), in which the diagonal matrix $\mathbf{D}$ is fixed as $\mathbf{D} = diag(1, 1, \ldots, 1)$, is also used for comparison. In addition, the miGraph

Table 1. Experiment results(%)

Algorithm	Precision	Recall	F-measure
CSC	81.62±0.72	83.38±0.87	82.46±0.19
SC	80.02±1.08	82.0±0.76	80.98±0.53
miGraph	80.01±1.59	80.82±0.92	80.4±1.06
MI-SVM	79.78	78.92	79.35
CKNN	78.85	70.54	74.46
EM-DD	77.59	72.97	75.21
SI-SVM	75.41	75.41	75.41

method [15] is also used to compare the performances between sparse coding and Support Vector Machines (SVM). The average accuracies of ten times 10-fold cross validation is used as the final performances for each method.

3.2. Results

For each data set, given the ground truth of a horror scene set (HS) as well as recognition results (ES) of an algorithm, the precision (P), recall (R), and F-measure (F_1) defined in Eq.(9) are used to evaluate the performances.

$$P = \frac{|HS \cap ES|}{|ES|}, R = \frac{|HS \cap ES|}{|HS|}, F_1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

The average Precision (P), Recall (R) and F-measure (F_1) are shown in Table 1. The methods MI-SVM, CKNN, EM-DD, SI-SVM in Table 1 denote the MIL based recognition methods with different MIL classifiers [12]. The results in Table 1 show that the performances of CSC, SC and miGraph methods that consider context cues inside a scene outperform the other MIL-based methods that consider the instances independently. The CSC has much higher value than SC. It indicates that the visual-audio context is much important for horror scene recognition; and the CSC model can effectively fuse the visual-audio features. In addition, the CSC has lower standard deviations that imply the stableness of CSC. Furthermore, the training free character embedded in the sparse coding classifier makes it possible to be extended as an online classifier that is necessary for many video analysis applications.

4. Conclusion

Most existing studies on horror scene recognition neglect the fact that two types of contexts, one among frames and the other one between visual cue and audio cue, play an important role in emotion expression recognition. In this paper, we have proposed a novel cost-sensitive sparse coding (CSC) model based on the graph kernel to model these two contextual relationships for the problem. The experimental results have shown that our model is superior to other existing horror detection methods.

Acknowldegement This work is partly supported by the National Nature Science Foundation of China (No. 61005030, 60935002 and 60825204) and Chinese National High-tech R&D Program (863 Program)(No.2012AA012503 and No. 2012AA012504).

References

[1] A. Hanjalic and L. Q. Xu. Affective video content representation and modeling. *IEEE TM*, 7(1):143–154, 2005.

[2] W. M. Hu, O. Wu, and Z. Chen. Recognition of pornographic web pages by classifying texts and images. *IEEE TPAMI*, 29(6):1019–1034, 2007.

[3] G. Irie, K. Hidaka, T. Satou, and et al. Latent topic driving model for movie affective scene classification. *ACM MM*, pages 565–568, 2009.

[4] H. B. Kang. Affective content detection using hmms. *ACM MM*, pages 259–262, 2003.

[5] H. Lee, A. Battle, R. Raina, and et al. Efficient sparse coding algorithms. *NIPS*, pages 359–367, 2006.

[6] B. Li, W. H. Xiong, and W. M. Hu. Context-aware multi-instance learning based on hierarchical sparse coding. *ICDM*, pages 370–377, 2011.

[7] B. Li, W. H. Xiong, and W. M. Hu. Web horror image recognition based on context-aware multi-instance learning. *ICDM*, pages 1158–1163, 2011.

[8] Z. Rasheed, Y. Sheikh, and M. Shah. On the use of computable features for film classification. *IEEE TCSVT*, 15(1):52–64, 2005.

[9] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[10] H. L. Wang and L. Cheong. Affective understanding in film. *IEEE TCSVT*, 16(6):689–704, 2006.

[11] J. C. Wang, B. Li, W. M. Hu, and et al. Horror movie scene recognition based on emotional perception. *ICIP*, pages 1489–1492, 2010.

[12] J. C. Wang, B. Li, W. M. Hu, and et al. Horror video scene recognition via mutiple-instance learning. *ICASSP*, pages 1325–1328, 2011.

[13] J. Wright, A. Y. Yang, A. Ganesh, and et al. Robust face recognition via sparse representation. *TPAMI*, pages 210–227, 2009.

[14] B. Wu, X. Jiang, T. Sun, and et al. A novel horror scene detection scheme on revised multiple instance learning model. *MMM*, pages 377–388, 2011.

[15] Z. Zhou, Y. Sun, and Y. Li. Multi-instance learning by treating instances as non-i.i.d. samples. *ICML*, pages 1249–1256, 2009.