Multi-Perspective Cost-Sensitive Context-Aware Multi-Instance Sparse Coding and Its Application to Sensitive Video Recognition[1]

Weiming Hu[1], Xinmiao Ding[2], Bing Li[1], Jianchao Wang[1], Yan Gao[1], Fangshi Wang[3], and Stephen Maybank[4]

wmhu@nlpr.ia.ac.cn; dingxinmiao@126.com; bli@nlpr.ia.ac.cn; jianchao1030@163.com; 870327884@qq.com; fshwang@bjtu.edu.cn; sjmaybank@dcs.bbk.ac.uk

[1]National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190

[2]Shandong Institute of Business and Technology

[3]Beijing Jiaotong University, Beijing 100044)

[4]Department of Computer Science and Information Systems, Birkbeck College, Malet Street, London WC1E 7HX

Abstract: With the development of video-sharing websites, P2P, micro-blog, mobile WAP websites, and so on, sensitive videos can be more easily accessed. Effective sensitive video recognition is necessary for web content security. Among web sensitive videos, this paper focuses on violent and horror videos. Based on color emotion and color harmony theories, we extract visual emotional features from videos. A video is viewed as a bag and each shot in the video is represented by a key frame which is treated as an instance in the bag. Then, we combine multi-instance learning (MIL) with sparse coding to recognize violent and horror videos. The resulting MIL-based model can be updated online to adapt to changing web environments. We propose a cost-sensitive context-aware multi-instance sparse coding (MI-SC) method, in which the contextual structure of the key frames is modeled using a graph, and fusion between audio and visual features is carried out by extending the classic sparse coding into cost-sensitive sparse coding. We then propose a multi-perspective multi-instance joint sparse coding (MI-J-SC) method that handles each bag of instances from an independent perspective, a contextual perspective, and a holistic perspective. The experiments demonstrate that the features with an emotional meaning are effective for violent and horror video recognition, and our cost-sensitive context-aware MI-SC and multi-perspective MI-J-SC methods outperform the traditional MIL methods and the traditional SVM and KNN-based methods.

Index terms: Cost-sensitive context-aware MI-SC, Multi-perspective MI-J-SC, Horror video recognition, Violent video recognition, and Video emotional feature extraction.

1. Introduction

The emergence and development of video-sharing websites, P2P, micro-blog, podcasting, mobile WAP websites, and 3GP websites facilitate the dissemination of sensitive videos, such as adult, horror, violent, and

terrorist videos. Fig. 1 shows some examples of violent videos and horror videos. Diffusion of sensitive videos poses a major threat to national security, social stability, and the physical, psychological, and mental health of viewers. Effective recognition of sensitive videos is necessary for web content security [54]. Recognition of sensitive videos is a newly emergent research topic in the multimedia and pattern recognition communities, in the context of multimedia retrieval [55, 56], multimedia content understanding [59, 60], and multimodal fusion [56, 57], etc. In recent years a number of specific attempts have been made to deal with the problem of sensitive video recognition, and most of them focus on adult video recognition [11, 12, 18, 21]. In this paper, we focus on recognition of horror videos and violent videos.

(a)

(b)

Fig. 1. Examples of frames taken from (a) violent videos and (b) horror videos.

1.1. Related work

Violent videos usually stimulate psychic impulses by showing the use of force to injure others or oneself. The contents of violent videos [51] include fights, gun shots, explosions, and self-mutilation. The current recognition methods usually use visual features or audio features separately or fuse visual and audio features. The visual features can be used to detect human violence, such as kicking and fist fighting, in videos [52]. For instance, Datta et al. [22] adopted an accelerated motion vector to detect fight scenes. Wang et al. [44] detected violence in videos using the accumulated squared derivative features which were extracted from dense trajectories derived from videos. Xu et al. [53] detected violent videos by capturing distinctive local shape and motion patterns. Audio features can be used to detect violent speech or actions. For instance, Cheng et al. [23] used a hierarchical audio-based method to identify car racing and gunplay. Theodoros et al. [24, 31] extracted eight audio features from the frequency and time domains to detect violent videos. Acar et al. [50] detected violent videos using mid-level audio features in a bag-of-audio words method using Mel-frequency

Cepstral coefficients (MFCCs). Visual and audio features can be combined to more accurately locate violent scenes. Nam et al. [32] recognized violent videos by detecting blood and flames and exploiting representative audio effects, such as explosions and gunshots. Smeaton et al. [35] combined visual and audio features to select representative shots in an action video. Giannakopoulos et al. [43] detected violence using the statistics of audio features and average motion and motion orientation variance features. Lin and Wang [45] combined auditory and visual classifiers in a co-training way to detect violent shots in movies.

Horror videos strive to elicit the primary emotions of fear, horror, and terror. The contents of horror videos include serial killings, ghosts, monsters, vampires, animal killing, and irreligion. Horror information may arouse fears in children and teenagers and even induce phobias [46, 47]. The earlier work [5, 6, 8] on horror video recognition was carried out as a part of a video scene classification based on human emotions. Specific work on horror video recognition with its own characteristics emerged [13, 14]. Xu et al. [14] detected audio emotional events to locate horror video segments in videos which are known a priori to contain such segments. Wu et al. [13] represented each video as a bag of independent frames and applied multi-instance learning (MIL) to horror video recognition.

The current methods for violent and horror video recognition have the following limitations:

- They focus on using low level visual, motion, and audio features, or they only use affective audio features. Research on affective color and visual semantics, together with affective audio semantics in violent and horror videos, is still exploratory, but the results of this research are available for application to violent and horror video recognition.
- The current methods only focus on independent frames and do not consider the underlying contextual cues within violent and horror videos, even though contextual cues between frames are useful for recognizing violence and horror.
- While contexts between frames are useful for recognizing violent and horror emotions, independent frame cues also have emotional content. The independent frame cues, contextual cues among frames, and holistic features of the entire video are different sources of information for violent and horror video recognition. Well-chosen features from different perspectives can embody a variety of discriminative information. The current violent and horror recognition algorithms do not include the fusion of multi-perspectives to improve their performance.
- Web information changes rapidly. The current violent and horror video algorithms, overall, are unable

to update the classifiers online when new training samples are obtained.

1.2. Our work

As a variant of supervised learning, each sample for multi-instance learning (MIL) is a bag of instances instead of a single instance. Each bag is given a discrete or real-valued label. In binary classification, a bag is considered as positive if at least one instance in it is positive, and considered as negative if all its instances are negative. As a prior, a violent or horror video contains at least one violent or horror shot[2], and all the shots in a non-violent or non-horror video are necessarily non-violent or non-horror. If a video is treated as a bag and a shot in the video is treated an instance in the bag, violent and horror video recognition is consistent with the framework of MIL. So, we use MIL to recognize violent and horror videos.

The most current models for MIL in common use, such as axis-parallel concepts [15], the diverse density (DD) method [25], the expectation-maximization version of diverse density (EM-DD) [27], the MI-kernel method [28], the mi-SVM and MI-SVM [19], the mi-Graph and MI-Graph [29], and the adaptive p-posterior mixture-model (PP-MM) kernel [42], are trained in batch settings, in which the entire training set is available before each training procedure begins. Babenko et al. [48] proposed an online MIL algorithm based on a boosting technique. However, this online method assumes that all the instances in a positive bag are positive. This assumption is easily violated in practical applications. Li et al [49] extended the MIL algorithm based on embedded instance selection [16, 17] to an online MIL algorithm. However, a classifier still needs to be retrained using the new samples. The citation-kNN [26] is not part of the training process. It determines the label of each test bag using the labeled bag samples nearest to the test bag and the bag samples whose nearest bag samples contain the test bag. However, the citation-kNN is sensitive to outlier samples. Sparse coding (SC) is training-free, and the model can be updated online each time the labeled sample set is updated. Furthermore, SC is not sensitive to outliers, because the sparsity regularization can suppress outliers in the sparse representation. Therefore, we combine MIL with sparse coding to form a multi-instance sparse coding (MI-SC) technique for recognizing violent and horror videos.

The contributions of our work are summarized as follows:

- We extract color emotional features according to the results from psychological experiments. These color emotional features bridge the affective semantic gap to some extent. The color emotional features together with low-level visual features, motion features, and audio features are used for

[2] A shot is a consecutive sequence of frames captured by a camera action which takes place between start and stop operations.

violent and horror video recognition.

- We propose a cost-sensitive context-aware MI-SC method which can make use of the context among frames in the same video and the context between visual and audio cues for violent and horror video recognition. A video is divided into a series of shots via shot segmentation and a key frame from each shot is selected. The visual feature vector of each key frame is extracted to represent the shot in which the key frame exists. An audio feature vector is extracted for the entire video. A video is represented as a bag of instances which correspond to the visual feature vectors. A graph is constructed using the key frames as nodes to represent their contextual relations. A cost-sensitive sparse coding model is constructed to represent the context between the bag of visual feature vectors and the audio feature vector. We solve the cost-sensitive context-aware MI-SC using the existing feature sign search algorithm via a mathematical transformation.
- We propose a multi-perspective multi-instance joint sparse coding (MI-J-SC) method to combine information from a contextual perspective, an independent perspective, and a holistic perspective. The contexts between key frames form only a contextual perspective for violent and horror video recognition. A key frame also includes semantic meaning, so treating a video as a bag of independent instances can be considered as an independent perspective. The holistic features for the entire video can be treated as another perspective. The information from different perspectives more fully describes a video. The current MIL lacks the ability to fuse multi-perspectives. We incorporate the joint sparse coding into multi-instance classification to fuse the features from multi-perspectives, in order to obtain more accurate recognition of violent and horror videos.

The experimental results show the effectiveness of the extracted video emotion features. The results on the violent and horror video datasets show that our methods outperform the traditional MIL-based methods and the traditional SVM and KNN-based methods. The results on the general MIL datasets show that our methods may be effective for other general multi-instance problems.

The remainder of this paper is organized as follows: Section 2 presents the MI-SC technique. Sections 3 and 4 propose our cost-sensitive context-aware MI-SC and multi-perspective MI-J-SC methods, respectively. Section 5 presents our method for extracting emotional features and our method for recognizing violent and horror videos. Section 6 reports the experimental results. Section 7 concludes this paper.

2. Multi-Instance Sparse Coding

Multi-instance sparse coding (MI-SC) carries out MIL using the sparse coding technique. In the following, we first briefly introduce sparse coding. Then, we describe the mechanism of MIL via sparse coding.

2.1. Sparse coding

The goal of sparse coding [20] is to represent each input vector approximately as a weighted linear combination of "basis vectors" such that a small number of weights are non-zero. Given an *h*-dimensional input vector $\mathbf{x} \in \mathbb{R}^h$ and *n* basis vectors $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n] \in \mathbb{R}^{h \times n}$, a sparse vector $\mathbf{w} \in \mathbb{R}^n$, whose entry w_j **(**$1 \le j \le n$**)** is the weight of $\mathbf{u}_j$**,** is found such that

$$\mathbf{x} \approx \mathbf{U}\mathbf{w} = \sum_{j=1}^{n} \mathbf{u}_j w_j \ . \tag{1}$$

The objective of sparse coding is usually formulated as the minimization of the reconstruction error with sparsity regularization:

$$\min_{\mathbf{w}} \left\| \mathbf{x} - \mathbf{U}\mathbf{w} \right\|_2^2 + \lambda \left\| \mathbf{w} \right\|_1 \tag{2}$$

where the ℓ_1 norm $\left\| \mathbf{w} \right\|_1$ of **w** is the sparsity term and λ is a regularization factor to control the sparsity of **w**.

2.2. MIL via sparse coding

For MIL, a training dataset $\{(\mathbf{X}_1, y_1), \cdots, (\mathbf{X}_i, y_i), \cdots, (\mathbf{X}_N, y_N)\}$ consists of N bags $\{\mathbf{X}_i\}_{i=1}^N$ and their labels $\{y_i\}_{i=1}^N$. A bag $\mathbf{X}_i$ consists of n_i instances: $\mathbf{X}_i = \{\mathbf{x}_{i,1}, \cdots, \mathbf{x}_{i,j}, \cdots, \mathbf{x}_{i,n_i}\}$, where each instance $\mathbf{x}_{i,j}$ is a vector. The task of MI-SC is to sparsely combine the training bags $\{\mathbf{X}_i\}_{i=1}^N$ to represent a test bag.

Due to the set structure of the bags, a test bag cannot directly be sparsely and linearly reconstructed using the training bags. We apply a mapping function $\tilde{\varphi}: \mathbf{X} \to \mathbb{R}^d$ to map each bag **X** to a high dimensional vector space: $\mathbf{X} \to \tilde{\varphi}(\mathbf{X})$ (the descriptions and handling of the mapping functions will be detailed in Section 3.3). Then, by mapping the training bags to the high dimensional vector space, a basis matrix $\mathbf{B} = [\tilde{\varphi}(\mathbf{X}_1), \tilde{\varphi}(\mathbf{X}_2), \cdots, \tilde{\varphi}(\mathbf{X}_N)]$ for sparse coding is obtained. Given a test bag $\mathbf{X}_t$, the sparse coding in the high dimensional vector space is defined as:

$$\min_{\mathbf{w}} \left\| \tilde{\varphi}(\mathbf{X}_t) - \mathbf{B}\mathbf{w} \right\|_2^2 + \lambda' \left\| \mathbf{w} \right\|_1 . \tag{3}$$

The label of $\mathbf{X}_t$ is determined by the labels of the training samples whose weights are nonzero for sparsely

representing $\mathbf{X}_t$. It is clear that this is a training-free online learning model which is updated only by changing the labeled samples. The limitation of the above MI-SC is that the contexts among instances are not modeled.

3. Cost-Sensitive Context-Aware MI-SC

To handle the above limitation, we formulate context-aware MI-SC and cost-sensitive sparse coding, and propose a method for optimizing the coefficients for the cost-sensitive context-aware MI-SC.

3.1. Context-aware MI-SC

Traditional MIL usually assumes that instances in a bag are independent of each other. Zhou et al. [29] built a graph [33] in their SVM-based MIL method to model the contexts between instances in each bag. This graph representation of contexts is incorporated into our MI-SC method.

For a bag $\mathbf{X}_i$, a graph G_i whose nodes are the instances in the bag is constructed. The distances between instances are computed. If the distance between two instances is smaller than a preset threshold, then the weight for the edge between the corresponding two nodes is set to 1, otherwise the weight is set to 0. A matrix $\mathbf{E}^i \in \mathbb{R}^{n_i \times n_i}$ of the adjacency weights for G_i is obtained, where $E^i_{a,a} = 1$ ($a = 1, 2, ..., n_i$).

The training samples are represented as $\{(\mathbf{X}_1, G_1, y_1), ..., (\mathbf{X}_i, G_i, y_i), ..., (\mathbf{X}_N, G_N, y_N)\}$, and a test bag is given as $(\mathbf{X}_t, G_t, y_t)$. We apply a mapping function $\varphi : G \to \mathbb{R}^d$ to map each graph G to a high dimensional vector space: $G \to \varphi(G)$. Then, the basis matrix for spare coding is replaced by $\mathbf{C} = [\varphi(G_1), \varphi(G_2), ..., \varphi(G_n)]$. The context-aware MI-SC is formulated as:

$$\min_{\mathbf{w}} \left\| \varphi(G_t) - \mathbf{C}\mathbf{w} \right\|_2^2 + \lambda'' \left\| \mathbf{w} \right\|_1 . \tag{4}$$

3.2. Cost-sensitive sparse coding

In real applications, each bag $\mathbf{X}_i$ may be associated with another kind of feature. For example, an audio is usually associated with a video, and the holistic features of the audio can overall characterize the entire video. We propose a cost-sensitive sparse representation to incorporate the associated features into the bags.

For each bag $\mathbf{X}_i$, its associated feature vector $\mathbf{a}_i$ is extracted. Then, the training set can be represented by $\{(\mathbf{a}_1, \mathbf{X}_1, G_1, y_1), (\mathbf{a}_2, \mathbf{X}_2, G_2, y_2), ..., (\mathbf{a}_N, \mathbf{X}_N, G_N, y_N)\}$. Given a test bag $\mathbf{X}_t$, we define a diagonal matrix $\mathbf{D} \in \mathbb{R}^{N \times N}$ whose diagonal entries are the Euclidean distances between the associated feature vector of the test bag and the associated feature vectors of each training bag:

$$\mathbf{D} = diag(\|\mathbf{a}_1 - \mathbf{a}_t\|, \cdots, \|\mathbf{a}_i - \mathbf{a}_t\|, \cdots, \|\mathbf{a}_N - \mathbf{a}_t\|) . \quad (5)$$

To incorporate the associated features into the MI-SC, we formulate cost-sensitive context-aware MI-SC in a high dimensional feature space as follows:

$$\min_{\mathbf{w}} \|\varphi(G_t) - \mathbf{C}\mathbf{w}\|_2^2 + \lambda''' \|\mathbf{D}\mathbf{w}\|_1 . \quad (6)$$

where the diagonal matrix **D** is included into the ℓ_1 norm in (4). The entries in **D** are cost values for the different training samples. In this way, the training samples, whose associated feature vectors have small distances to the associated feature vector of the test bag, are more likely to be selected to reconstruct the test bag. In the sensitive video recognition application, the videos which have audio tracks similar to the test video are more likely to be chosen to represent the test video.

3.3. Optimization

The traditional sparse coding optimization methods cannot be directly applied to the cost-sensitive context-aware MI-SC in (6). We transform the objective function in (6) to a form to which the traditional sparse coding optimization can be applied. Then the feature sign search (FSS) algorithm is used to solve for the coefficient vector **w**. Let $\mathbf{q} = \mathbf{D}\mathbf{w}$, where $\mathbf{q} \in \mathbb{R}^N$. In order to ensure that **D** is invertible, we add a very small value ε to the diagonal entries of **D**, and obtain an inverse as follows:

$$\mathbf{D}^{-1} = diag\left(\left(\|\mathbf{a}_1 - \mathbf{a}_t\| + \varepsilon\right)^{-1}, \left(\|\mathbf{a}_2 - \mathbf{a}_t\| + \varepsilon\right)^{-1}, ..., \left(\|\mathbf{a}_N - \mathbf{a}_t\| + \varepsilon\right)^{-1} \right) \quad (7)$$

Substituting $\mathbf{w}=\mathbf{D}^{-1}\mathbf{q}$ into (6) yields:

$$\min_{\mathbf{q}} \|\varphi(G_t) - \mathbf{C}\mathbf{D}^{-1}\mathbf{q}\|_2^2 + \lambda''' \|\mathbf{q}\|_1 . \quad (8)$$

Let $\mathbf{V} = \mathbf{C}\mathbf{D}^{-1}$, where $\mathbf{V} \in \mathbb{R}^{d \times N}$. Formula (8) is rewritten as:

$$\min_{\mathbf{q}} \|\varphi(G_t) - \mathbf{V}\mathbf{q}\|_2^2 + \lambda''' \|\mathbf{q}\|_1 . \quad (9)$$

The function $\varphi(\cdot)$ which is used to map bags into a high dimensional space is difficult to define explicitly. Instead, the scalar product $\varphi(G_i)^T \varphi(G_j)$ in the high dimensional space is explicitly defined via a kernel function. So, we transform the objective of (9) into a form involving scalar products $\varphi(G_i)^T \varphi(G_j)$. It is clear that

$$\|\varphi(G_t) - \mathbf{V}\mathbf{q}\|_2^2 = [\varphi(G_t) - \mathbf{V}\mathbf{q}]^T [\varphi(G_t) - \mathbf{V}\mathbf{q}] = [\varphi(G_t)]^T \varphi(G_t) + \mathbf{q}^T \mathbf{V}^T \mathbf{V}\mathbf{q} - 2\mathbf{q}^T \mathbf{V}^T \varphi(G_t) . \quad (10)$$

Then, we only need to consider $\mathbf{V}^T\mathbf{V}$ and $\mathbf{V}^T\varphi(G_t)$. It is clear that

$$\begin{aligned}\mathbf{V}^T\mathbf{V} &= \left(\mathbf{C}\mathbf{D}^{-1}\right)^T\left(\mathbf{C}\mathbf{D}^{-1}\right) = (\mathbf{D}^{-1})^T\mathbf{C}^T\mathbf{C}\mathbf{D}^{-1} \\ &= (\mathbf{D}^{-1})^T\left[\varphi(G_1),\varphi(G_2),...,\varphi(G_N)\right]^T\left[\varphi(G_1),\varphi(G_2),...,\varphi(G_N)\right]\mathbf{D}^{-1} \\ &= (\mathbf{D}^{-1})^T\begin{bmatrix}\varphi(G_1)^T\varphi(G_1) & \varphi(G_1)^T\varphi(G_2) & ... & \varphi(G_1)^T\varphi(G_N) \\ \varphi(G_2)^T\varphi(G_1) & \varphi(G_2)^T\varphi(G_2) & ... & \varphi(G_2)^T\varphi(G_N) \\ ... & ... & ... & ... \\ \varphi(G_N)^T\varphi(G_1) & \varphi(G_N)^T\varphi(G_2) & ... & \varphi(G_N)^T\varphi(G_N)\end{bmatrix}\mathbf{D}^{-1}\end{aligned} \quad (11)$$

and

$$\begin{aligned}\mathbf{V}^T\varphi(G_t) &= (\mathbf{C}\mathbf{D}^{-1})^T\varphi(G_t) = (\mathbf{D}^{-1})^T[\varphi(G_1),\varphi(G_2),...,\varphi(G_N)]^T\varphi(G_t) \\ &= (\mathbf{D}^{-1})^T\begin{bmatrix}\varphi(G_1)^T\varphi(G_t) \\ \varphi(G_2)^T\varphi(G_t) \\ \cdots \\ \varphi(G_N)^T\varphi(G_t)\end{bmatrix}\end{aligned} \quad . \quad (12)$$

It remains to define a graph kernel function $K_g()$ to represent the scalar product $\varphi(G_i)^T\varphi(G_j)$ of graphs G_i and G_j in the high dimensional feature space. The definition of a graph kernel function depends on a kernel function between any two instances. The Gaussian radial basis function (RBF) kernel $K(\mathbf{x}_{i,a},\mathbf{x}_{j,b})$ between an instance $\mathbf{x}_{i,a}$ in bag *i* and an instance $\mathbf{x}_{j,b}$ in bag *j* is defined as:

$$K(\mathbf{x}_{i,a},\mathbf{x}_{j,b}) = \exp\left(-\rho\left\|\mathbf{x}_{i,a}-\mathbf{x}_{j,b}\right\|_2^2\right) \quad (13)$$

where ρ is a scaling factor. Let $\omega_{i,a}$ be the weight for the instance $\mathbf{x}_{i,a}$, in bag *i*, which is defined as:

$$\omega_{i,a} = \frac{1}{\sum_{u=1}^{n_i} E_{a,u}^i} \quad (14)$$

where *u* is the index for an instance in bag *i* and $\mathbf{E}^i$ is the adjacency weight matrix for bag *i*. The kernel function $\mathrm{K}_g()$ [29] between graphs G_i and G_j is defined as:

$$\mathrm{K}_g(G_i,G_j) = \frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}\omega_{i,a}\omega_{j,b}K(\mathbf{x}_{i,a},\mathbf{x}_{j,b})}{\sum_{a=1}^{n_i}\omega_{i,a}\sum_{b=1}^{n_i}\omega_{j,b}} \quad . \quad (15)$$

Using the graph kernel function, the objective function in (9) is explicitly formulated, and then the optimization in (9) is efficiently solved by the recently proposed feature-sign search algorithm (FSS) [30].

3.4. Classification

After the optimal coefficient vector **q** is obtained, we calculate the reconstruction residual of the test bag for each bag label *m*, and the label with the smallest reconstruction residual is selected as the label to which

the test bag belongs. For each label *m*, we define a vector $\boldsymbol{\delta}^m \in \mathbb{R}^N$ whose *l*-th entry δ_l^m is:

$$\delta_l^m = \begin{cases} q_l, & y_l = m \\ 0, & y_l \neq m \end{cases} \tag{16}$$

i.e., this vector only selects coefficients associated with labels *m*. The reconstruction residual $\mho_m(G_t)$ of the test bag for label *m* is defined as:

$$\mho_m(G_t) = \left\| \varphi(G_t) - \mathbf{V}\boldsymbol{\delta}^m \right\|_2^2 = 1 + (\boldsymbol{\delta}^m)^T \mathbf{V}^T \mathbf{V} \boldsymbol{\delta}^m - 2(\boldsymbol{\delta}^m)^T \mathbf{V}^T \varphi(G_t) \tag{17}$$

where $\varphi(G_t)^T \varphi(G_t) = 1$. We assign the test bag the final label *c* which is defined as follows:

$$c = \arg_m \left(\min(\mho_m(G_t)) \right). \tag{18}$$

4. Multi-Perspective Multi-Instance Joint Sparse Coding

Based on the structured joint sparse representation [2, 3, 34], we propose multi-perspective cost-sensitive MI-J-SC which includes the above cost-sensitive context-aware MI-SC.

4.1. Structured joint sparse representation

It is assumed that there are *K* different types of feature and *M* labels in the training dataset. Let $\boldsymbol{\Psi}_m^k \in \mathbb{R}^{h_k \times N_m}$ be the matrix of each feature *k* (*k*=1,2,..,*K*) for the training samples with label *m*, where h_k is the dimension of the *k*-th type of feature and N_m is the number of the training samples with label *m*: $\sum_{m=1}^{M} N_m = N$. Then, the matrix of the *k*-th type of feature for all the training samples is $\boldsymbol{\Psi}^k = [\boldsymbol{\Psi}_1^k, \boldsymbol{\Psi}_2^k, ..., \boldsymbol{\Psi}_M^k]$. The *k*-th type's feature vector $\mathbf{z}^k \in \mathbb{R}^{h_k}$ of a test sample $\mathbf{Z}$ is reconstructed from the *k*-th feature vectors of the training samples:

$$\mathbf{z}^k = \sum_{m=1}^{M} \boldsymbol{\Psi}_m^k \mathbf{w}_m^k + \boldsymbol{\varepsilon}^k \tag{19}$$

where $\mathbf{w}_m^k \in \mathbb{R}^{N_m}$ is the reconstruction coefficient vector for the *k*-th feature vectors of the samples with label *m*, and $\boldsymbol{\varepsilon}^k$ is the residual term. Let $\mathbf{w}^k = \left[(\mathbf{w}_1^k)^T, (\mathbf{w}_2^k)^T, ..., (\mathbf{w}_M^k)^T \right]^T \in \mathbb{R}^N$ be the coefficient vector for the *k*-th type of feature. Let $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^K] \in \mathbb{R}^{N \times K}$. The $\ell_{2,1}$-mixed norm of $\mathbf{W}$ is:

$$\|\mathbf{W}\|_{2,1} = \sum_{m=1}^{M} \sqrt{\sum_{k=1}^{K} \left\| \mathbf{w}_m^k \right\|_2^2} = \sum_{m=1}^{M} \left\| \mathbf{W}_m \right\|_2 \tag{20}$$

where $\mathbf{W}_m = [\mathbf{w}_m^1, \mathbf{w}_m^2, ... \mathbf{w}_m^K] \in \mathbb{R}^{N_m \times K}$. Then, the reconstruction in (19) can be represented by the least square

regression based on the $\ell_{2,1}$ mixed- norm regularization [2, 3, 34]:

$$\min_{\mathbf{W}}\left(\frac{1}{2}\sum_{k=1}^{K}\left\|\mathbf{z}^k-\sum_{m=1}^{M}\mathbf{\Psi}_m^k\mathbf{w}_m^k\right\|_2^2+\lambda\left\|\mathbf{W}\right\|_{2,1}\right). \tag{21}$$

The $\ell_{2,1}$ mixed-norm includes the ℓ_2 norm of the vector of the coefficients of the *K* feature vectors for each training sample and the ℓ_1 norm of the vector of the ℓ_2 norm values for all the samples. The $\ell_{2,1}$ mixed-norm guarantees joint sparse representation. The reasons are summarized as follows:

- The ℓ_1 norm in the $\ell_{2,1}$ mixed-norm ensures that the training samples chosen to represent a test sample are as few as possible.
- The ℓ_2 norm in the $\ell_{2,1}$ norm ensures that when a training sample is not chosen to represent a test sample, all the *K* feature vectors of the training sample are not chosen to represent the test sample.

This structured joint sparse coding can effectively fuse information from multiple features.

4.2. Multi-perspectives of multi-instances

We extend the above structured joint sparse representation to MIL to fuse information from multi-perspectives. Different perspectives can be defined according to different applications. We define the following three multi-perspectives in the context of sensitive video recognition:

1) Independent perspective: As in traditional MIL, the instances in a bag are treated as independent. We define a mapping function $\varphi^1:\mathbf{X}\to\mathbb{R}^{d_1}$ to map the feature space of the bags $\{\mathbf{X}\}$ to a d_1-dimensional vector space: $\mathbf{X}\to\varphi^1(\mathbf{X})$. Then, the training samples are transformed to $\mathbf{F}^1=[\varphi^1(\mathbf{X}_1),\cdots,\varphi^1(\mathbf{X}_i),\cdots,\varphi^1(\mathbf{X}_N)]$. In the d_1-dimensional vector space, we define a kernel function $\mathrm{K}_1()$ between any two bags $\mathbf{X}_i$ and $\mathbf{X}_j$ as follows:

$$\mathrm{K}_1(\mathbf{X}_i,\mathbf{X}_j)=[\varphi^1(\mathbf{X}_i)]^T\varphi^1(\mathbf{X}_j)=\frac{\sum_{a=1}^{n_i}\sum_{b=1}^{n_j}K(\mathbf{x}_{ia},\mathbf{x}_{jb})}{\sum_{l=1}^{n_i}\sum_{l'=1}^{n_i}K(\mathbf{x}_{il},\mathbf{x}_{il'})\sum_{s=1}^{n_j}\sum_{s'=1}^{n_j}K(\mathbf{x}_{js},\mathbf{x}_{js'})} \tag{22}$$

where the kernel *K*() between two instances is defined as in (13).

2) Contextual perspective: The graph constrained MI-SC in Section 3.1 is introduced to form a contextual perspective for MIL. We define a mapping function $\varphi^2:G\to\mathbb{R}^{d_2}$ to map the features of each bag with a graph *G* to a d_2-dimensional space: $G\to\varphi^2(G)$ (φ^2 is just φ in (4)). The context-aware training

bags are transformed to $\mathbf{F}^2 = [\varphi^2(G_1), \cdots, \varphi^2(G_i), \cdots, \varphi^2(G_N)]$. In the d_2-dimensional vector space, the kernel function $\mathrm{K}_2()$ is defined as in (15).

3) Holistic perspective: Statistical histograms of instances in bags can be used for bag classification. From a holistic perspective, we construct a feature histogram for a bag based on the bag-of-words model [4]. Given the set of the training bag samples, all the instances are clustered to form a lexicon of *R* code words $\{\mathbf{d}_1, \cdots, \mathbf{d}_r, ..., \mathbf{d}_R\}$. Each instance $\mathbf{x}_{ij}$ in a bag $\mathbf{X}_i$ is mapped to a code word $\pi(\mathbf{x}_{ij})$ which is determined by:

$$\pi(\mathbf{x}_{ij}) = \arg\min_{1 \le r \le R} \left\| \mathbf{x}_{ij} - \mathbf{d}_r \right\|. \tag{23}$$

In bag $\mathbf{X}_i$, the number of occurrences $h(r, \mathbf{X}_i)$ of each code word r ($r = 1, 2, ..., R$) is counted: $h(r, \mathbf{X}_i) = \left| \{\mathbf{x}_{ij} \in \mathbf{X}_i : \pi(\mathbf{x}_{ij}) = r\} \right|$, where $|\cdot|$ is the number of entries in a set. Then, bag $\mathbf{X}_i$ is represented by a normalized histogram ξ_i:

$$\xi_i = \left[\frac{h(1, \mathbf{X}_i)}{\sum_{r=1}^{R} h(r, \mathbf{X}_i)}, \cdots, \frac{h(r, \mathbf{X}_i)}{\sum_{r=1}^{R} h(r, \mathbf{X}_i)}, \cdots, \frac{h(R, \mathbf{X}_i)}{\sum_{r=1}^{R} h(r, \mathbf{X}_i)} \right]. \tag{24}$$

Then, the set of the training samples is represented by $\{(\mathbf{X}_1, \xi_1, y_1), \cdots, (\mathbf{X}_i, \xi_i, y_i), \cdots, (\mathbf{X}_N, \xi_N, y_N)\}$. We map each histogram feature vector to a high dimensional feature space using a mapping function $\varphi^3 : \xi_i \to \mathbb{R}^{d_3}$. Then, the histograms of the training samples are transformed to $\mathbf{F}^3 = [\varphi^3(\xi_1), ..., \varphi^3(\xi_i), ..., \varphi^3(\xi_N)]$. In this high dimensional space, we define the kernel function between any two bags as follows:

$$\mathrm{K}_3(\xi_i, \xi_j) = [\varphi^3(\xi_i)]^T \varphi^3(\xi_j) = \sum_{r_1=1}^{R} \sum_{r_2=1}^{R} \xi_i[r_1] \xi_j[r_2] K(\mathbf{d}_{r_1}, \mathbf{d}_{r_2}) \tag{25}$$

where $K(\mathbf{d}_{r_1}, \mathbf{d}_{r_2})$ is the Gaussian kernel function between two code words $\mathbf{d}_{r_1}$ and $\mathbf{d}_{r_2}$:

$$K(\mathbf{d}_{r_1}, \mathbf{d}_{r_2}) = \exp(-\sigma \left\| \mathbf{d}_{r_1} - \mathbf{d}_{r_2} \right\|^2). \tag{26}$$

4.3. Multi-perspective cost-sensitive MI-J-SC

We use the structured joint sparse representation in Section 4.1 to fuse the information from multi-perspectives such as defined in Section 4.2. Also, cost-sensitive sparse coding can be applied to the structured joint sparse representation. Then, we propose a multi-perspective cost-aware MIL method by integrating multi-perspectives into a unified joint sparse coding framework based on the $\ell_{2,1}$ norm. Given *K* perspectives (*K* is 3 in this paper), the training sample set is represented by *K* matrices $\{\mathbf{F}^1, \mathbf{F}^2, ..., \mathbf{F}^K\}$, where

$\mathbf{F}^k = [\varphi^k(\mathbf{X}_1), \varphi^k(\mathbf{X}_2), ..., \varphi^k(\mathbf{X}_N)]$. Given a test sample $\mathbf{X}_t$, its feature vector in each perspective *k* is represented by $\mathbf{f}^k = \varphi^k(\mathbf{X}_t)$. Let $\mathbf{w}^k \in \mathbb{R}^N$ be the coefficient vector for the training samples at perspective *k* and $\mathbf{W}$ be the matrix of the coefficient vectors of the *K* perspectives: $\mathbf{W} = [\mathbf{w}^1, \mathbf{w}^2, ..., \mathbf{w}^K] \in \mathbb{R}^{N \times K}$. Then, multi-perspective cost-sensitive MI-SC is represented by:

$$\min_{\mathbf{W}} \left(\frac{1}{2} \sum_{k=1}^{K} \left\| \mathbf{f}_t^k - \mathbf{F}^k \mathbf{w}^k \right\|_2^2 + \eta \left\| \mathbf{D}\mathbf{W} \right\|_{2,1} \right) \tag{27}$$

where $\mathbf{D}$ is the cost matrix defined in (5). In (27), the first term is the sum of the squared reconstruction errors from different perspectives, and the second term is the regularization to control the sparsity of the coefficients. We group the training feature set $\mathbf{F}^k$ of each perspective *k* according to the class labels $\{m\}_{m=1}^M$ of the training samples: $\mathbf{F}^k = [\mathbf{F}_1^k, ..., \mathbf{F}_m^k, ..., \mathbf{F}_M^k]$ where $\mathbf{F}_m^k$ is the matrix which consists of the *k*-th feature vectors of the training samples with label *m*. Accordingly, the *k*-th coefficient vector in $\mathbf{W}$ is also grouped as: $\left[(\mathbf{w}_1^k)^T, ..., (\mathbf{w}_m^k)^T, ..., (\mathbf{w}_M^k)^T \right]^T$. Let $\mathbf{W}_m = [\mathbf{w}_m^1, \mathbf{w}_m^2, ..., \mathbf{w}_m^K] \in \mathbb{R}^{N_m \times K}$ (*m*=1,2,..,*M*), where N_m is the number of the training samples with class *m*. Then, Equation (27) is rewritten as:

$$\min_{\mathbf{W}} \left(\sum_{k=1}^{K} \frac{1}{2} \left\| \mathbf{f}_t^k - \sum_{m=1}^{M} \mathbf{F}_m^k \mathbf{w}_m^k \right\|_2^2 + \eta \sum_{m=1}^{M} \left\| \mathbf{D}_m \mathbf{W}_m \right\|_{2,1} \right) \tag{28}$$

where $\mathbf{D}_m \in \mathbb{R}^{N_m \times N_m}$ is the diagonal matrix whose entries are those elements in $\mathbf{D}$ corresponding to the training samples with label *m*.

4.4. Optimization

The $\ell_{2,1}$ mixed-norm accelerated proximal gradient (APG) algorithm [34] is introduced to optimize the object function in (28). The APG cannot be directly applied to (28). We make a transformation to (28). Let $\mathbf{Q}_m = \mathbf{D}_m \mathbf{W}_m$ where $\mathbf{Q}_m = [\mathbf{q}_m^1, \mathbf{q}_m^2, ..., \mathbf{q}_m^K] \in \mathbb{R}^{N_m \times K}$. $\mathbf{W}_m = \mathbf{D}_m^{-1} \mathbf{Q}_m$. $\mathbf{F}_m^k \mathbf{w}_m^k = \mathbf{F}_m^k (\mathbf{D}_m)^{-1} \mathbf{q}_m^k$. Let $\mathbf{U}_m^k = \mathbf{F}_m^k (\mathbf{D}_m)^{-1}$. It follows that (28) is equivalent to:

$$\min_{\mathbf{W}} \left(\sum_{k=1}^{K} \frac{1}{2} \left\| \mathbf{f}_t^k - \sum_{m=1}^{M} \mathbf{U}_m^k \mathbf{q}_m^k \right\|_2^2 + \eta \sum_{m=1}^{M} \left\| \mathbf{Q}_m \right\|_2 \right). \tag{29}$$

The APG algorithm can be applied to (29).

The APG algorithm alternately updates a coefficient matrix $\mathbf{Q}^t = [\mathbf{q}_m^{k,t}]$ and an aggregation matrix $\mathbf{V}^t = [\mathbf{v}_m^{k,t}]$ at each iteration *t* which consists of a generalized gradient mapping step and an aggregation step.

In the generalized gradient mapping step, given the current aggregation matrix $\hat{\mathbf{V}}^t$, the coefficient matrix $\mathbf{Q}^t$ is updated. Let $\mathbf{U}^k = [\mathbf{U}_1^k, ..., \mathbf{U}_m^k, ..., \mathbf{U}_M^k]$. It is clear that

$$\mathbf{U}_k^T \mathbf{U}_k = (\mathbf{D}^{-1})^T \begin{bmatrix} \varphi^k(\mathbf{X}_1)^T \varphi^k(\mathbf{X}_1) & \varphi^k(\mathbf{X}_1)^T \varphi^k(\mathbf{X}_2) & ... & \varphi^k(\mathbf{X}_1)^T \varphi^k(\mathbf{X}_N) \\ \varphi^k(\mathbf{X}_2)^T \varphi^k(\mathbf{X}_1) & \varphi^k(\mathbf{X}_2)^T \varphi^k(\mathbf{X}_2) & ... & \varphi^k(\mathbf{X}_2)^T \varphi^k(\mathbf{X}_N) \\ ... & ... & ... & \cdots \\ \varphi^k(\mathbf{X}_N)^T \varphi^k(\mathbf{X}_1) & \varphi^k(\mathbf{X}_N)^T \varphi^k(\mathbf{X}_2) & ... & \varphi^k(\mathbf{X}_N)^T \varphi^k(\mathbf{X}_N) \end{bmatrix} \mathbf{D}^{-1} \tag{30}$$

and

$$\mathbf{U}_k{}^T \varphi^k(\mathbf{X}_t) = (\mathbf{D}^{-1})^T \begin{bmatrix} \varphi^k(\mathbf{X}_1)^T \varphi^k(\mathbf{X}_t) \\ \varphi^k(\mathbf{X}_2)^T \varphi^k(\mathbf{X}_t) \\ ... \\ \varphi^k(\mathbf{X}_N)^T \varphi^k(\mathbf{X}_t) \end{bmatrix} \tag{31}$$

where the scalar product $\varphi^k(\mathbf{X}_i)^T \varphi^k(\mathbf{X}_j)$ between bags $\mathbf{X}_i$ and $\mathbf{X}_j$ is evaluated using a kernel function which is explicitly defined in (15), (22), or (25). A matrix $\mathbf{P}^t = [\mathbf{p}^{1,t}, ..., \mathbf{p}^{k,t}, ..., \mathbf{p}^{K,t}] \in \mathbb{R}^{N \times K}$ is defined as follows:

$$\mathbf{p}^{k,t} \leftarrow -\mathbf{U}_k{}^T \varphi(\mathbf{X}_t) + \mathbf{U}_k{}^T \mathbf{U}_k v^{k,t}, \ k=1,2,..,K. \tag{32}$$

Then,

$$\mathbf{q}^{k,t+1} \leftarrow \mathbf{v}^{k,t} - \mu \mathbf{q}^{k,t}, \ k=1,2,\ldots,K \tag{33}$$

and

$$\mathbf{q}_m^{t+1} \leftarrow \max\left(1 - \frac{\eta\mu}{\left\|\mathbf{q}_m^{t+1}\right\|_2}, 0\right) \mathbf{q}_m^{t+1}, \ \ m=1,\cdots,M \tag{34}$$

where μ is the step size parameter.

In the aggregation step, the aggregation matrix is updated by constructing a linear combination of $\mathbf{Q}^t$ and $\mathbf{Q}^{t+1}$:

$$\mathbf{V}^{t+1} \leftarrow \mathbf{Q}^{t+1} + \frac{\tau_{t+1}(1-\tau_t)}{\tau_t}(\mathbf{Q}^{t+1} - \mathbf{Q}^t) \tag{35}$$

where conventionally $\tau_t = 2/(t+2)$ [36].

4.5. Classification

Using the obtained optimal coefficient matrix **Q**, the reconstruction residual $\mathfrak{O}_m(\mathbf{X}_t)$ of the test bag for label $m \in \{1,\cdots,M\}$ is defined as:

$$\mho_m(\mathbf{X}_t)=\sum_{k=1}^{K}\left\|\varphi^k(\mathbf{X}_t)-\mathbf{U}_m^k\mathbf{q}_m^k\right\|_2^2=K+\sum_{k=1}^{K}\left((\boldsymbol{\delta}_m(\mathbf{q}^k))^T(\mathbf{U}^k)^T\mathbf{U}^k\boldsymbol{\delta}_m(\mathbf{q}^k)-2(\mathbf{U}^k)^T\varphi^k(\mathbf{X}_t)\boldsymbol{\delta}_m(\mathbf{q}^k)\right) \quad (36)$$

where $\boldsymbol{\delta}_m(\mathbf{q}^k)$ is a coefficient selector that only selects coefficients associated with label *m*, i.e., the *l*-th entry in $\boldsymbol{\delta}_m(\mathbf{q}^k)$ is defined as follows:

$$d_m(\mathbf{q}^k)_l=\begin{cases}q_l, & y_l=m\\ 0, & y_l\neq m\end{cases}. \quad (37)$$

Similar to (18), the label that has the smallest residual is assigned to the test bag $\mathbf{X}_t$.

5. Sensitive Video Recognition

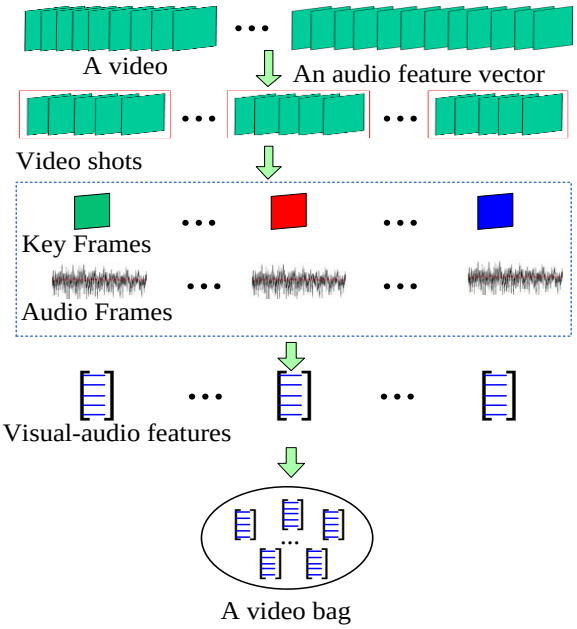

Fig. 2. Bag construction for each video.

We apply the proposed cost-sensitive context-aware MI-SC and multi-perspective MI-J-SC to recognize sensitive videos, especially horror videos and violent videos. Given a set of *N* videos $\{\mathbf{I}_1,\mathbf{I}_2,...,\mathbf{I}_N\}$, they are labeled as $\{y_1,y_2,...,y_N\}$ ($y_i\in\{1,2\}$, i.e., *M*=2) where a sensitive video is labeled "1" and a non-sensitive video is labeled as "2". Each video $\mathbf{I}_i$ is divided into n_i shots $\{\mathbf{s}_{i,1},\mathbf{s}_{i,2},...,\mathbf{s}_{i,n_i}\}$ by measuring mutual information and joint entropy between frames [37]. In each shot, we select the frame which is closest to the mean of the color emotional features in the shot as a key frame, and then a key frame set $\{\boldsymbol{\theta}_{i,1},\boldsymbol{\theta}_{i,2},...,\boldsymbol{\theta}_{i,n_i}\}$ for video $\mathbf{I}_i$ is obtained. The visual and audio feature vector $\mathbf{f}_{i,j}$ for each key frame $\boldsymbol{\theta}_{i,j}$ is extracted. An audio feature vector $\mathbf{a}_i$ is extracted from the entire audio associated with $\mathbf{I}_i$. A bag for each video is constructed by treating the feature vector of each key frame as an instance, as shown in Fig. 2. Then, the above MI-SC

methods can be applied to $\{\mathbf{I}_1, \mathbf{I}_2, ..., \mathbf{I}_N\}$. The optimal coefficients obtained by the cost-sensitive context-aware MI-SC or the multi-perspective MI-J-SC are used to classify the test videos as sensitive or non-sensitive. In the following, we describe the features extracted from horror and violent videos.

The features extracted from videos are based on emotional perception theory. Different colors, textures, and audio rhythms may produce different emotions. So, we extract the following video features that produce emotions in the viewers: color emotional features, visual features, and audio features. These emotion-producing features are used for horror video recognition. Additional motion features are used for recognizing violent videos.

5.1. Color emotional features

Ou et al. [38, 39] developed color emotion models for single colors and harmony models for two color combinations by psychophysical experiments. We extract color emotional features based on these color emotion models.

Ou et al. found that color emotions for single-colors depend on the following three factors: activity, weight, and heat, which are defined as follows:

$$\begin{cases} Activity = -2.1 + 0.06\sqrt{(L^* - 50)^2 + (a^* - 3)^2 + ((b^* - 17)/1.4)^2} \\ Weight = -1.8 + 0.04(100 - L^*) + 0.45\cos(h - 100^0) \\ Heat = -0.5 + 0.02(C^*)^{1.07}\cos(h - 50^0) \end{cases} \quad (38)$$

where (L^*, a^*, b^*) and (L^*, C^*, h) are the color components in the CIELAB and CIELCH color spaces, respectively. Based on (38), we define an emotional intensity (EI) for each pixel (*x*,*y*) as follows:

$$EI(x, y) = \sqrt{Activity^2 + Weight^2 + Heat^2}\ . \quad (39)$$

Given a frame in a video, the EIs for all the pixels are computed. Based on the EIs, a color emotion histogram is acquired and employed as part of the color emotional features.

Ou and Luo [1] developed a quantitative two-color harmony model which consists of three independent color harmony factors: hue effect (H_H), lightness effect (H_L), and chromatic effect (H_C). These three harmony factors for two colors are explicitly estimated using hues, saturations, and lightness values of these two colors in the CIELAB color space (The details can be found in [1]). The overall harmony score *CH* between these two colors is defined as the sum of the three factors: $CH = H_H + H_C + H_L$. Given a frame, for each pixel we calculate the color harmony score CH_1 between its color and the mean of the colors of its surrounding pixels and the color harmony score CH_2 between its color and the mean of the colors of all the

pixels in the frame. The color harmony score CH_f of this pixel is defined as the sum of the two scores: $CH_f = CH_1 + CH_2$. Based on the color harmony scores in the frame, we construct a color harmony histogram which is used as another part of the emotional features.

5.2. Visual features

The visual emotional features include lighting features, color features, texture features, and Rhythm features.

1) Lighting feature: Lighting affects viewers' feelings directly [5, 7]. The lighting effect is determined by two factors: the general level of light and the proportion of shadow area. We use the median of the L values of all the pixels in a frame in the Luv color space [7] to characterize the general level of the light in the frame. The proportion of the pixels, whose lightness values are below a certain shadow threshold, is used to estimate the proportion of shadow area.

2) Color feature: The color values used in the HSV space are clearly distinguishable by human perception, so we use the means and variances of components of the HSV color space in a frame to characterize the main cues of colors in the frame. Particular colors have strong relations with movie genres [7]. The particular colors in a frame can be represented by the covariance matrix $\mathbf{\Theta}$ of the L, u, v values of pixels in the frame:

$$\mathbf{\Theta} = \begin{bmatrix} \sigma_L^2 & \sigma_{L,u}^2 & \sigma_{L,v}^2 \\ \sigma_{L,u}^2 & \sigma_L^2 & \sigma_{u,v}^2 \\ \sigma_{L,v}^2 & \sigma_{u,v}^2 & \sigma_v^2 \end{bmatrix}. \tag{40}$$

The determinant of (40), $\Sigma = \det(\mathbf{\Theta})$, is used as the feature for the particular colors.

3) Texture feature: Texture is another important factor relevant to image emotion, because different textures give people different feelings. Geusebroek and Smeulders [40] proposed a six-stimulus basis for stochastic texture perception: Texture distributions in image scenes conform to a Weibull distribution associated with a random variable x:

$$wb(x) = \frac{\gamma}{\beta}\left(\frac{x}{\beta}\right)^{\gamma-1} e^{-\frac{1}{\gamma}\left(\frac{x}{\beta}\right)^{\gamma}} \tag{41}$$

where β represents the contrast of the image (a higher value for β indicates more contrast), and γ represents the grain size of the image (a higher value for γ indicates a smaller grain size, i.e., more fine textures). The parameters β and γ completely characterize the spatial structure of the texture, and they are used as the texture feature for horror and violent video recognition.

4) Rhythm feature: In horror and violent videos, quick shot switching and strong motions are often used to excite nervous moods in the viewers. We use the inverse length of a shot to represent the speed of shot switching. For a frame, the mean and standard deviation of motions between frames in a short clip centered at the frame are used to measure the quantity of motion associated with the frame [10].

5.3. Audio features

Specific sounds and music are often used to highlight emotional atmosphere and promote dramatic effects. The following audio features [9] are extracted:

- The mean and variance of the 12 MFCCs (Mel-frequency Cepstral coefficients) of each frame and the 12 MFCCS' first-order differential, where the MFCCs are computed from the fast Fourier transform (FFT) power coefficients.
- Spectral power which is used to measure the energy intensity of an audio signal: For an audio signal $s(t)$, each frame is weighted with a Hamming window $h(t)$, where t is the index of a sample in the frame. The spectral power of an audio frame of the signal $s(t)$ is calculated as:

$$10\log\left[\frac{1}{T}\left\|\sum_{t=0}^{T-1}s(t)h(t)\exp\left(-j2\pi\frac{to}{T}\right)\right\|^2\right] \tag{42}$$

  where T is the number of samples of each frame, and o is the index of an order of the DFT coefficients.
- The mean and variance of the spectral centroids of the audio signal, which are employed as measures of music brightness.
- Time domain zero crossings rate which provides a measure of the noisiness of an audio signal.

5.4. Motion features

The following motion features are extracted especially for violent video recognition:

1) Optical flow: Corners in the previous frame are detected. Optical flow is used to estimate the positions of the corners in the current frame. The distance moved by each corner between the previous and current frames is calculated. The sum, mean, and standard deviation of the distances moved by the corners are used as motion features.

2) Motion template: The motion template is constructed using the motion history image obtained from consecutive frames. The motion template is segmented into a number of regions. The global motion orientation of the template and the mean and standard deviation of the motion orientations of all these regions

are calculated and used as additional motion features.

Empirically, all the above color emotional features, visual features, and audio features are useful for both horror video recognition and violent video recognition. The above motion features are useful only for violent video recognition, rather than horror video recognition, because violent videos use much more intense motions than horror videos to trigger strong emotions. In each instance the used features are combined into a feature vector. The value of each component in a feature vector is normalized according to the maximum of the values of this component over all the samples in the dataset. For the cost-sensitive context-aware MI-SC, all the audio features extracted from an entire video form a single audio feature vector for the video. This audio feature vector is used to calculate the cost matrix. For the multi-perspective MI-J-SC, the same features are used in all the three perspectives.

6. Experiments

In the experiments, the color emotion histogram has 64 bins. The color harmony histogram has 25 bins. The shadow threshold for lighting feature extraction was experimentally determined as 0.18. For an audio signal, we extracted a single-channel audio stream at 44.1 KHz and computed 12 MFCCs over 20ms frames.

We used the precision (P), recall (R), and F1-measure (F_1) to evaluate the performance of an algorithm. Let HS be the horror or violent videos in a dataset, and ES be the videos that are recognized as horror or violent by the algorithm. The precision (P), recall (R), and F1-measure (F_1) are defined as:

$$\begin{cases} P = \dfrac{|HS \cap ES|}{|ES|} \\ R = \dfrac{|HS \cap ES|}{|HS|} \\ F_1 = \dfrac{2 \times P \times R}{P + R} \end{cases} \tag{43}$$

The proposed sensitive video recognition methods were compared with the following methods:

- **EM-DD** [27]: This is an MIL method which combines the EM algorithm with the diverse density (DD) maximization [25].
- **mi-Graph** [29]: This method uses a graph [33] to model the contexts between instances in a bag.
- **MI-kernel** [28]: This method regards each bag as a set of feature vectors and then applies a set-based kernel directly for bag classification.
- **MI-SVM**: This method is extended from SVM to deal with MIL problems. It represents a positive bag by the instance farthest from the separating hyper-plane.

- **mi-SVM**: It looks for the hyper-plane such that for each positive bag there is at least one instance lying in the positive half-space, and all the instances belonging to negative bags lie in the negative half-space.
- **Citation-KNN**: It **is** extended from KNN to deal with MIL problems. It considers not only the labels of the bags which are nearest to the test bag, but also the labels of the bags whose nearest samples contain the test bag.
- **SVM**: The feature vectors of the key frames in a video were averaged into one vector. These averaged feature vectors of all the training samples were used to construct a classical SVM-based sensitive video classifier.
- **KNN:** The KNN, instead of the SVM in the SVM-based classifier, was used to train a classifier.

In the following, we report first the results of horror video recognition, then the results of violent video recognition, and finally the results on the general MIL datasets for validating the effect of proposed MI-SC methods.

6.1. Horror video recognition

We downloaded horror and non-horror videos from the internet. This dataset consists of 400 horror videos and 400 non-horror videos. These videos come from different countries, such as China, US, Japan, South Korea, and Thailand. The genres of the non-horror movies include comedy, action, drama, and cartoon. Half of the horror videos and half of the non-horror videos were used for training, and the remaining videos were used for testing. The average accuracies of ten times 10-fold cross validation were used to measure the performance of each method.

Table 1 shows the values of the average Precision (P), Recall (R) and F1 measure (F_1) of our methods based on cost-sensitive context-aware MI-SC and multi-perspective cost-sensitive context-aware MI-J-SC, and also the values for the competing methods based on mi-Graph, MI-kernel, MI-SVM, Citation-KNN, EM-DD, SVM, and KNN. In order to validate the effectiveness of the audio cost in the method based on cost-sensitive context-aware MI-SC, the audio features were not included in the feature vector in each instance when testing the method based on cost-sensitive context-aware MI-SC. We also compared the method based on cost-sensitive context-aware MI-SC with the method based on the pure context-aware MI-SC obtained by removing the audio cost from this method, i.e., the diagonal matrix D was fixed as D=diag(1 , 1, . . . , 1). It is unfair to compare the method based on the pure context-aware MI-SC with the

competing methods in which the audio features are used. Therefore, we removed the audio features from the feature vectors and compared the method based on the pure context-aware MI-SC with the mi-Graph-based method without audio features. From Table 1, the following points were revealed:

Table 1. The experimental results on the horror video dataset (%)

Algorithms	Precision (*P*)	Recall (*R*)	F1 measure
Multi-perspective cost-sensitive MI-J-SC	85.56(±0.51)	85.21(±0.39)	85.38(±0.32)
Cost-sensitive context-aware MI-SC	81.62(±0.72)	83.38(±0.87)	82.46(±0.19)
Pure context-aware MI-SC	80.02(±1.08)	82.00(±0.76)	80.98(±0.53)
miGraph with audio features	81.87(±1.95)	82.4(±1.25)	82.14(±1.20)
miGraph without audio features	80.01(±1.59)	80.82(±0.92)	80.40(±1.06)
MI-kernel	80.70(±1.42)	81.43(±0.9)	81.05(±0.5)
MI-SVM	79.78	78.92	79.35
Citation-KNN	78.85	70.54	74.46
EM-DD	77.59	72.97	75.21
SVM	75.41	75.41	75.41
kNN	89.10	57.30	69.70

- Our method based on multi-perspective cost-sensitive MI-J-SC is much more accurate than all the other methods. This shows that horror video recognition benefits from multi-perspectives. The lower standard deviations imply that our method is stable.
- The method based on the cost-sensitive context-aware MI-SC has a higher mean F1 value and a much lower standard deviation than the method based on the pure context-aware MI-SC. This indicates that the visual-audio context is useful for horror video recognition and the method based on the cost-sensitive MI-SC effectively fuses the visual and audio features.
- Our method based on cost-sensitive context-aware MI-SC, our method based on multi-perspective cost-sensitive context-aware MI-J-SC, the method based on the pure context-aware MI-SC and the method based on the mi-Graph method, all of which model contextual cues among instances in a bag, outperform other MIL-based methods in which the instances are treated independently. This shows that the contextual relations between instances are useful for horror video recognition.
- The results of the mi-Graph-based MIL method, in which SVMs rather than sparse coding are used, are reported. It is seen that our method based on cost-sensitive context-aware MI-SC yields more accurate results than the mi-Graph-based method with audio features. Although the method based on the pure context-aware MI-SC yields less accurate results than the mi-Graph-based method with audio features, it yields more accurate results than the mi-Graph-based method without audio

features. It is apparent that the sparse coding-based MIL methods outperform the SVM-based MIL method for horror video recognition.

- The two non-MIL-based methods, the KNN-based method and the pure SVM-based method, overall yield less accurate results than the MIL-based methods. This is because they use holistic features in videos. If a horror video contains only a small number of horror frames, then the holistic features inevitably weaken the features obtained from the horror frames. The pure SVM-based method outperforms the KNN-based method, because SVM considers experiential risk and structural risk.

Furthermore, the training free characteristic **of** the sparse coding classifiers makes it feasible to extend our methods based on cost-sensitive context-aware MI-SC and multi-perspective cost-sensitive context-aware MI-J-SC to online classifiers that are necessary for network video analysis applications.

The computational efficiency of the proposed model is ensured by the efficient optimization methods for obtaining the sparsity coefficient vector. The feature sign search (FSS) algorithm in the cost-sensitive context-aware MI-SC produces a significant speedup for sparse coding. The APG algorithm in the multi-perspective cost-sensitive MI-J-SC is a fast algorithm for solving the $\ell_{2,1}$ norm-regularized optimization. Table 2 compares the runtimes of the proposed methods and other representative methods on the horror video dataset tested on a computer with Intel(R) Core(TM)2 Quad CPU. It is seen that the test speed of our sparse coding-based methods is comparable to other representative methods, not taking into account the fact that our methods have no training time.

Table 2. Runtime in seconds per video for different methods

	Training	Test
Multi-perspective cost-sensitive MI-J-SC	0	0.07
Cost-sensitive context-aware MI-SC	0	0.05
mi-Graph	1.02	0.05
mi-kernel	1.02	0.06
SVM	1.06	0.04
Citation-kNN	0	0.19

In the experiments we fused different features to show their different contributions. Seven different combinations of the visual features (VF), the audio features (AF), and the color emotion features (EF) were obtained. Table 3 shows the precision, recall, and F1 measure for multi-perspective MI-J-SC, mi-Graph, MI-SVM, SVM, and kNN using these seven feature combinations on the horror video dataset. It is seen that the best one among three types of features is the audio feature, which has the highest F1 measure. Generally,

the combination of the visual features, the audio features, and the color emotional features can improve the recognition accuracy, which shows the complementary characteristics of the three types of features.

Table 3. The results for different feature combinations (%)

		VF	EF	AF	VF+EF	VF+AF	EF+AF	VF+EF+AF
Multi-perspective MI-J-SC	Precision	70.1	69.8	81.0	68.8	81.6	84.1	85.4
	Recall	72.7	73.6	81.8	70.4	81.3	84.7	85.2
	F1 measure	71.4	71.7	81.4	69.6	81.5	84.4	85.3
mi-Graph	Precision	69.1	69.9	80.8	76.4	82.7	82.7	84.0
	Recall	68.8	68.5	81.3	77.0	81.5	84.8	83.0
	F1 measure	68.9	69.2	81.0	76.7	81.8	83.7	83.5
MI-SVM	Precision	72.2	73.3	71.3	72.0	84.1	81.0	80.7
	Recall	73.3	74.3	81.7	74.0	83.3	81.8	82.8
	F1 measure	72.8	73.8	81.5	73.0	83.7	81.4	81.8
SVM	Precision	72.8	70.6	76.9	71.0	77.4	75.6	78.6
	Recall	73.0	70.3	76.5	71.5	77.0	76.5	79.0
	F1 measure	72.9	70.5	76.7	71.3	77.2	76.1	78.8
kNN	Precision	76.6	72.9	86.0	74.5	88.0	89.4	89.1
	Recall	58.0	71.3	50.5	57.0	49.5	57.0	57.3
	F1 measure	66.0	72.1	63.6	64.6	63.4	69.6	69.7

6.2. Violent video recognition

Table 4. The experimental results on the violent video dataset (%)

Algorithms	Precision (P)	Recall (R)	F1 measure
Multi-perspective cost-sensitive MI-J-SC	86.57(±0.48)	87.87(±0.87)	87.2(±0.53)
Cost-sensitive context-aware MI-SC	86.13(±0.98)	85.95(±0.95)	86.04(±0.87)
mi-Graph	85.95(±2.28)	85.82(±1.69)	85.85(±1.15)
MI-kernel	84.77(±3.37)	84.99(±2.84)	84.79(±1.25)
MI-SVM	82.75	82.54	82.64
Citation-KNN	78.88	83.33	81.04
EM-DD	71.01	84.52	77.17
SVM	82.25	79.66	80.93
kNN	80.78	73.59	77.02

We downloaded violent and non-violent movies from the internet. This dataset consists of 400 violent videos and 400 non-violent videos. Half of the violent videos and half of the non-violent videos were used for training, and the remaining videos were used for testing. The average accuracies of ten times 10-fold cross validation were used as the final performances for each method. Table 4 shows the recognition results of our methods based on cost-sensitive context-aware MI-SC and multi-perspective cost-sensitive MI-J-SC, and the competing methods based on mi-Graph, MI-kernel, MI-SVM, Citation-KNN, EM-DD, SVM, and KNN. All the methods use the same features including color emotional features, visual features, audio features, and motion features which are all introduced in Section 5. It is seen that our methods yield more accurate results than the competing methods, and our multi-perspective cost-sensitive context-aware MI-J-SC method yields

more accurate results than our cost-sensitive context-aware MI-SC method. The results have the same characteristics as on the horror dataset.

We also tested performance of violent video recognition methods on the VSD (violent scene detection) 2014 dataset [62], which benchmarks violence detection in Hollywood movies at the MediaEval benchmarking initiative for multimedia evaluation. The training set in the dataset has 24 Hollywood movies and contains binary annotations of all the violent scenes, where a scene was identified by its start and end frames. A set of 7 Hollywood movies was used for testing. All the test violent segments were annotated at video frame level, i.e., a violent segment was defined by its starting and ending frame numbers. We segmented the test videos into scenes and labeled the scenes as violent or non-violent using the videos' annotations at the frame level. Table 5 compares the results of our methods for detecting violent scenes with the state-of-the-art results on the dataset. It is seen that the results of our methods are better than the stat-of-the-art results. The effectiveness of the extracted features and the MI-SC-based classification in our methods is clearly shown.

Table 5. Comparison between the results (%) of our methods and the state-of-the-art results on the Hollywood movie test set and the YouTube movie test set, respectively

Test subset	Method	Precision	Recall	F1 measure
Hollywood movies	Multi-perspective MI-J-SC	63.8	69.8	66.6
	Context-aware MI-SC	50.8	69.5	58.7
	FUDAN [63]	41.1	72.1	52.4
	RECOD [61]	33.0	69.7	44.8
	VIVOLAB [58]	38.1	58.4	46.1

6.3. MIL datasets

Although we focused our multi-perspective context-aware MI-J-SC method on applications to sensitive video recognition, our method can be used in other applications. To verify the generality of our multi-perspective context-aware MI-J-SC method, we tested it on the general datasets which were widely used to evaluate the performance of MIL methods. They include five benchmark datasets**:** Musk1, Musk2, Elephant, Fox, and Tiger [15, 19]. The Musk1 and Musk2 datasets are musk molecule datasets. Each molecule which corresponds to a bag has several shape structures which correspond to instances. Each structure was represented by a 166 dimensional vector. The Musk1 dataset contains 47 positive and 45 negative bags. The Musk2 dataset contains 39 positive and 63 negative bags. The Elephant, Fox and Tiger datasets are image datasets. Each image which corresponds to a bag was segmented into several image patches which correspond to instances. A 230 dimensional vector was extracted from each patch. Each of these three datasets contains

100 positive and 100 negative bags.

Table 6. Accuracy (%) on the MIL benchmark datasets

Algorithm	Musk1	Musk2	Elephant	Fox	Tiger
Multi-perspective MI-J-SC	91.1(±2.8)	90.6(±1.3)	88.5(±1.1)	62.7(±1.8)	86.8(±1.2)
mi-Graph	88.9(±3.3)	90.3(±2.6)	86.8(±0.7)	61.6(±2.8)	86.0(±1.6)
MI-Graph	90.0(±3.8)	90.0(±2.7)	85.1(±2.8)	61.2(±1.7)	81.9(±1.5)
MI-Kernel	88.0(±3.1)	89.3(±1.5)	84.3(±1.6)	60.3(±1.9)	84.2(±1.0)
MI-SVM	77.9	84.3	81.4	59.4	84.0
mi-SVM	87.4	83.6	82.0	58.2	78.9
Miss-SVM	87.6	80.0	N/A	N/A	N/A
PP-MM	95.6	81.2	82.4	60.3	82.4
DD	88.0	84.0	N/A	N/A	N/A
EM-DD	84.8	84.9	78.3	56.1	72.1

We compared our multi-perspective context-aware MI-J-SC method with the methods based on mi-Graph, MI-Graph, MI-Kernel, MI-SVM, mi-SVM [19], Miss-SVM [41], PP-MM kernel [42], the diverse density (DD) [25], and EM-DD [27]. For all the methods the same features from the benchmark datasets were used. The performance of each method was evaluated using the accuracy which is the proportion of the samples which are correctly classified. Our multi-perspective context-aware MI-J-SC method and the methods based on mi-Graph, MI-Graph, and MI-Kernel were run by us. The 10-fold cross validations for ten times were carried out to yield the average accuracies and standard deviations. The results of the competing methods based on MI-SVM and mi-SVM [19], Miss-SVM [41], PP-MM kernel [42], DD [25], and EM-DD [27] were directly taken from [29]. All the results are shown in Table 6. It is seen that our multi-perspective MI-J-SC method achieves better performances than the methods based on MI-Graph and mi-Graph on the Musk1, Elephant and Fox datasets. The performances of the methods based on multi-perspective MI-J-SC, MI-Graph, mi-Graph, and MI-Kernel on the Musk2 and Tiger datasets are comparable. More importantly, our multi-perspective MI-J-SC method yields lower standard deviations on all the benchmark datasets. This shows the stability of our multi-perspective context-aware MI-J-SC method.

7. Conclusion

In this paper, we have proposed a cost-sensitive context-aware MI-SC method in which a graph kernel has been used to model the contexts among frames and cost-sensitive sparse coding has been used to model the contexts between visual cues and audio cues. We have also proposed a multi-perspective MI-SC method which can effectively fuse information from the contextual perspective, the independent instance perspective, and the holistic perspective. Based on the color emotion and color harmony theories, we have extracted each

video's color emotional features which are higher level features in contrast with the low-level color and visual features. These color emotional features together with the cost-sensitive context-aware MI-SC method and the multi-perspective MI-J-SC method have been applied to recognize violent and horror videos. Experimental results have shown that the extracted emotional features are effective for recognizing violent and horror videos. It has been shown that our methods not only are superior to traditional MIL-based methods and traditional SVM and KNN-based methods on the violent and horror video datasets but also may be effective in other general multi-instance problems as tested on the general MIL datasets. Although this paper focuses on the recognition of violent and horror videos, our cost-sensitive context-aware MI-SC method and our multi-perspective MI-J-SC method are available for recognizing other types of web videos.

References

1. L.C. Ou and M.R. Luo, "A colour harmony model for two-colour combinations," *Color Research & Application*, vol. 31, no. 3, pp. 191-204, 2006.
2. F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint l2,1-norms minimization," *in Proc. of Advances in Neural Information Processing Systems*, pp. 1813-1821, 2010.
3. J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2,1-norm minimization," *in Proc. of Conference on Uncertainty in Artificial Intelligence*, pp. 339-348, 2009.
4. J. Winn, A. Criminisi, and T. Minka, "Object categorization by learned universal visual dictionary," in *Proc. of IEEE International Conference on Computer Vision*, vol. 2, pp. 1800-1807, 2005.
5. H.L. Wang and L. Cheong, "Affective understanding in film," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 16, no. 6, pp.689-704, 2006.
6. A. Hanjalic and L.Q. Xu, "Affective video content representation and modeling," *IEEE Trans. on Multimedia*, vol. 7, no. 1, pp.143-154, 2005.
7. Z. Rasheed, Y. Sheikh, and M. Shah, "On the use of computable features for film classification," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 15, no. 1, pp. 52-64, 2005.
8. H.B. Kang, "Affective content detection using HMMs," *in Proc. of ACM international conference on Multimedia*, pp. 259-262, 2003.
9. G. Tzanetakis and P. Cook, "Musical genre classification of audio signals," *IEEE Trans. on speech and audio processing*, vol. 10, no. 5, pp. 293-302, 2002.
10. S. Zhu and K.-K. Ma, "A new diamond search algorithm for fast block-matching motion estimation," *IEEE Trans. on Image Processing*, vol. 9, no. 2, pp. 287-290, Feb. 2000.
11. Y. Liu, X. Wang, Y. Zhang, and S. Tang, "Fusing audio-words with visual features for pornographic video detection," *in Proc. of IEEE International Conference on Trust, Security and Privacy in Computing and Communications*, pp. 1488-1493, 2011.
12. C. Jansohn, A. Ulges, and T.M. Breuel, "Detecting pornographic video content by combining image features with motion information," *in Proc. of ACM international conference on Multimedia*, Beijing, pp. 601-604, 2009.
13. B. Wu, X. Jiang, T. Sun, S. Zhang, X. Chu, C. Shen, and J. Fan. "A novel horror scene detection scheme on revised multiple instance learning model," *in Proc. of International Conference on Advances in Multimedia Modeling*, pp. 377-388, 2011.
14. M. Xu, L.-T. Chia, and J. Jin, "Affective content analysis in comedy and horror videos by audio emotional event detection," *in Proc. of IEEE International Conference on Multimedia and Expo*, pp. 622-625, July 2005.

15. T.G. Dietterich, R.H. Lathrop, and T. Lozano-Perez, "Solving the multiple-instance problem with axis-parallel rectangles," *Artificial Intelligence*, vol. 89, no. 1-2, pp. 31-71, 1997.
16. Y. Chen, J. Bi, and J.Z. Wang, "MILES: Multiple-instance learning via embedded instance selection," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1931-1947, 2006.
17. Y. Chen and J.Z. Wang, "Image categorization by learning and reasoning with regions," *Journal of Machine Learning Research*, vol. 5, pp. 913-939, 2004.
18. S. Lee, W. Shim, and S. Kim, "Hierarchical system for objectionable video detection," *IEEE Trans. on Consumer Electronics*, vol. 55, no. 2, pp. 677-684, May 2009.
19. S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support vector machines for multiple instance learning," *in Proc. of Advances in Neural Information Processing Systems*, pp. 561-568, Cambridge, MIT Press, 2003.
20. J. Wright, A. Yang, A. Ganesh, S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210-227, Feb. 2009.
21. T. Endeshaw, J. Garcia, and A. Jakobsson, "Classification of indecent videos by low complexity repetitive motion detection," *in Proc. of IEEE Applied Imagery Pattern Recognition Workshop*, Washington DC, pp. 1-7, Oct. 2008.
22. A. Datta, M. Shah, and N.D.V. Lobo, "Person-on person violence detection in video data," *in Proc. of International Conference on Pattern Recognition*, pp. 433-438, 2002.
23. W.-H. Cheng, W.-T. Chu, and J.-L. Wu, "Semantic context detection based on hierarchical audio models," *in Proc. of ACM SIGMM International Workshop on Multimedia Information Retrieval*, pp. 109-115, 2003.
24. T. Giannakopoulos, D. Kosmopoulos, A. Aristidou, and S. Theodoridis, "Violence content classification using audio features," *Advances in Artificial Intelligence, Lecture Notes in Computer Science*, vol. 3955, pp. 502-507, 2006.
25. O. Maron and T. Lozano-P ́erez, "A framework for multiple-instance learning," *in Proc. of Advances in Neural Information Processing Systems*, pp. 570-576. Cambridge, MIT Press, 1998.
26. J. Wang and J.-D. Zucker, "Solving the multi-instance problem: a lazy learning approach," *in Proc. of International Conference on Machine Learning*, pp. 1119-1125, 2000.
27. Q. Zhang and S.A. Goldman, "EM-DD: an improved multi-instance learning technique," *in Proc. of Advances in Neural Information Processing Systems*, Cambridge, MIT Press, pp. 1073-1080, 2001.
28. T. Gartner, P.A. Flach, A. Kowalczyk, and A.J. Smola, "Multi-instance kernels," *in Proc. of International Conference on Machine Learning*, pp. 179-186, 2002.
29. Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-instance learning by treating instances as non-I.I.D. samples," *in Proc. of International Conference on Machine Learning*, pp. 1249-1256, 2009.
30. H. Lee, A. Battle, R. Raina, and Y.N. Andrew, "Efficient sparse coding algorithms," *in Proc. of Advances in Neural Information Processing Systems*, pp. 359-367, 2006.
31. T. Giannakopoulos, A. Pikrakis, and S. Theodoridis, "A multi-class audio classification method with respect to violent content in movies using Bayesian networks," *in Proc. of IEEE Workshop on Multimedia Signal Processing*, pp. 90-93, Oct. 2007.
32. J. Nam, M. Alghoniemy, and A.H. Tewfik, "Audio-visual content-based violent scene characterization," *in Proc. of IEEE International Conference on Image Processing*, pp. 353-357, 1998.
33. J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319-2323, 2000.
34. X.T. Yuan and S. Yan, "Visual classification with multi-task joint sparse representation," *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3493-3500, June 2010.
35. A.F. Smeaton, B. Lehane, N.E. O'Connor, C. Brady, and G. Craig, "Automatically selecting shots for action movie trailers," *in Proc. of ACM International Workshop on Multimedia Information Retrieval*, pp. 231-238, 2006.
36. P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," *SIAM Journal of Optimization*, May 2008. http://www.math.washington.edu/~tseng/papers/apgm.pdf
37. Z. Cernekova, I. Pitas, and C. Nikou, "Information theory-based shot cut/fade detection and video summarization," *IEEE*

Trans. on circuits and systems for video technology, vol.16, no. 1, pp. 82-91, 2006.

38. L.C. Ou, M.R. Luo, A. Woodcock, and A. Wright, "A study of colour emotion and colour preference. part I: colour emotions for single colours," *Color Research & Application*, vol. 29, no. 3, pp. 232-240, 2004.
39. L.C. Ou, M.R. Luo, A. Woodcock, and A. Wright, "A study of colour emotion and colour preference. part III: colour preference modeling," *Color Research & Application*, vol. 29, no. 5, pp. 381-389, 2004.
40. J.M. Geusebroek and A.W.M. Smeulders, "A six-stimulus theory for stochastic texture," *International Journal of Computer Vision*, vol. 62, no. 1, pp. 7-16, 2005.
41. Z.-H. Zhou and J.-M. Xu, "On the relation between multi-instance learning and semi-supervised learning," *in Proc. of International Conference on Machine Learning*, pp. 1167-1174, 2007.
42. H.Y. Wang, Q. Yang, and H. Zha, "Adaptive p-posterior mixture model kernels for multiple instance learning," *in Proc. of International Conference on Machine Learning*, pp. 1136-1143, 2008.
43. T. Giannakopoulos, A. Makris, D. Kosmopoulos, S. Perantonis, and S. Theodoridis, "Audio-visual fusion for detecting violent scenes in videos," *Artificial Intelligence: Theories, Models and Applications, Lecture Notes in Computer Science*, vol. 6040, pp. 91-100, 2010.
44. K. Wang, Z. Zhang, and L. Wang, "Violence video detection by discriminative slow feature analysis," *in Proc. of Chinese Conference on Pattern Recognition*, pp. 137-144, Sep. 2012.
45. J. Lin and W. Wang, "Weakly-supervised violence detection in movies with audio and video based co-training," *Advances in Multimedia Information Processing, Lecture Notes in Computer Science*, vol. 5879, pp. 930-935, 2009.
46. A. Field and J. Lawson, "Fear information and the development of fears during childhood: effects on implicit fear responses and behavioural avoidance," *Behaviour Research and Therapy*, vol. 41, no. 11, pp. 1277-1293, Nov. 2003.
47. N.J. King, G. Eleonora, and T.H. Ollendick, "Etiology of childhood phobias: current status of Rachman's three pathways theory," *Behaviour Research and Therapy*, vol. 36, no. 3, pp. 297-309, 1998.
48. B. Babenko, M.-H. Yang, and S. Belongie, "Visual tracking with online multiple instance learning," *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 983-990, 2009.
49. M. Li, J. Kwok, and B.L. Lu, "Online multiple instance learning with no regret," *in Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1395-1401, 2010.
50. E. Acar, F. Hopfgartner, and S. Albayrak, "Detecting violent content in Hollywood movies by mid-level audio representations," *in Proc. of International Workshop on Content-Based Multimedia Indexing*, pp. 73-78, June 2013.
51. X. Ding, B. Li, W. Hu, W. Xiong, and Z. Wang, "Horror video scene recognition based on multi-view multi-instance learning," *in Proc. of Asian Conference on Computer Vision*, pp. 599-610, 2012.
52. H.-D. Kim, S.-S. Ahn, K.-H. Kim, and J.-S. Choi, "Single-channel particular voice activity detection for monitoring the violence situations," *in Proc. of IEEE International Symposium on Robot and Human Interactive Communication*, Korea, pp. 412-417, Aug. 2013.
53. L. Xu, C. Gong, J. Yang, Q. Wu, and L. Yao, "Violent video detection based on MoSIFT feature and sparse coding," *in Proc. of IEEE International Conference on Acoustic, Speech and Signal Processing*, pp. 3538-3542, 2014.
54. P.Y. Lee, S.C. Hui, and A.C.M. Fong, "An intelligent categorization engine for bilingual web content filtering", *IEEE Trans. on Multimedia*, vol. 7, no. 6, pp. 1183-1190, 2005.
55. M. Soleymani, M. Larson, T. Pun, and A. Hanjalic, "Corpus development for affective video indexing," *IEEE Trans. on Multimedia*, vol. 16, no. 4, pp. 1075-1089, 2014.
56. C. Kang, S. Xiang, S. Liao, C. Xu, and C. Pan, "Learning consistent feature representation for cross-modal multimedia retrieval," *IEEE Trans. on Multimedia*, vol. 17, no. 3 , pp. 370-381, 2015.
57. J. Geng, Z. Miao, and X. Zhang, "Efficient heuristic methods for multimodal fusion and concept fusion in video concept detection," *IEEE Trans. on Multimedia*, vol. 17, no. 4, pp. 498-511, 2015.
58. D. Castan, M. Rodriguez, A. Ortega, C. Orrite, and E. Lleida, "ViVoLab and CVLab-MediaEval 2014: violent scenes detection affect task," *in Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, Oct. 2014.

59. C. Tekin and M. van der Schaar, "Contextual online learning for multimedia content aggregation," *IEEE Trans. on Multimedia*, vol.17, no. 4, pp. 549-561, 2015.
60. Z. Ma, Y. Yang, N. Sebe, and K. Zheng, and A.G. Hauptmann, "Multimedia event detection using a classifier-specific intermediate representation," *IEEE Trans. on Multimedia*, vol. 15 , no. 7, pp. 1628-1637, 2013.
61. S. Avila, D. Moreira, M. Perez, D. Moraes, I. Cota, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, "RECOD at MediaEval 2014: violent scenes detection task," *in Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, Oct. 2014.
62. M. Schedl, M. Sjobergy, I. Mironicaz, B. Ionescuz, and V.L. Quangx, Y.-G. Jiang, and C.-H. Demartyk, "VSD2014: a dataset for violent scenes detection in Hollywood movies and web videos," *in Proc. of International Workshop on Content-Based Multimedia Indexing*, pp. 1-6, June 2015.
63. Q. Dai, Z. Wu, Y.-G. Jiang, X. Xue, and J. Tang, "Fudan-NJUST at MediaEval 2014: violent scenes detection using deep neural networks," *in Working Notes Proc. MediaEval 2014 Workshop*, Barcelona, Catalunya, Spain, Oct. 2014.

Acknowledgment

This work is partly supported by the 973 basic research program of China (Grant No. 2014CB349303), the Natural Science Foundation of China (Grant No. 61472421, 61303086), the Project Supported by CAS Center for Excellence in Brain Science and Intelligence Technology, and the Project Supported by Guangdong Natural Science Foundation (Grant No. S2012020011081)

Weiming Hu received the Ph.D. degree from the department of computer science and engineering, Zhejiang University in 1998. From April 1998 to March 2000, he was a postdoctoral research fellow with the Institute of Computer Science and Technology, Peking University. Now he is a professor in the Institute of Automation, Chinese Academy of Sciences. His research interests are in visual motion analysis, recognition of web objectionable information, and network intrusion detection.

Xinmiao Ding received the M.S. degree in electronic engineering from Dalian Maritime University, Dalian, China, in 2004, and the Ph. D. degree from the School of Mechanical Electronic and Information Engineering, China University of Mining and Technology, Beijing, China, in 2013. She is currently a visiting scholar in the Institute of Automation, Chinese Academy of Sciences. Her main research interests include Image and video analysis and understanding, machine learning and internet security.

Bing Li received the PhD degree from the Department of Computer Science and Engineering, Beijing Jiaotong University, China, in 2009. Currently, he is an associate professor in the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences. His research interests include color constancy, visual saliency and web content mining.

Jianchao Wang got the bachelor degree from the University of Science and Technology of China in 2008 and the master degree from the National Laboratory of Pattern Recognition in 2011. He is a an image processing engineer in the company of meituan. His research interests include image processing and computer vision.

Yan Gao received the B.S. degree in electrical engineering from the North University of China, Taiyuan, in 2013. He is currently pursuing the M.S. degree in the Civil Aviation University of China, Tianjin. His research interests include object recognition, image detection, and image retrieval.

Fangshi Wang is a professor with Software Engineering School in Beijing Jiaotong University. She received the PhD degree from the School of Computer Science and Engineering, Beijing Jiaotong University, China, in 2007. Her research interests focus on Computer vision, Video analysis and semantic tag.

Stephen Maybank received a BA in Mathematics from King's college Cambridge in 1976 and a PhD in computer science from Birkbeck college, University of London in 1988. Now he is a professor in the School of Computer Science and Information Systems, Birkbeck College. His research interests include the geometry of multiple images, camera calibration, visual surveillance etc.