# Horror Video Scene Recognition based on Multi-view Multi-instance Learning

Anonymous ACCV 2012 submission

Paper ID 393

**Abstract.** Comparing with the research of pornographic content filtering on Web, Web horror content filtering, especially horror video scene recognition is still on the stage of exploration. Most existing methods identify horror scene only from independent frames, ignoring the context cues among frames in a video scene. In this paper, we propose a Multi-view Multi-Instance Leaning (M$^2$IL) model based on joint sparse coding technique that takes the bag of instances from independent view and contextual view into account simultaneously and apply it on horror scene recognition. Experiments on a horror video dataset collected from internet demonstrate that our method's performance is superior to the other existing algorithms.

## 1 Introduction

Along with the rapid growing of the Internet, more and more information sources and services are available on the Web everyday, including pornography, violence, horror information, etc., which are not appropriate for all users, especially children. To protect our psychological health, effective content-filtering systems that can automatically block all objectionable contents are necessary for Web content security. Lots of scientific researchers have investigated into this area. Some of these system, for example pornographic content filters, have matured to a point where robust recognition or filtering software is available [1]. However, the research on affective semantics of horror video scene is still on the stage of exploration. Therefore, an effective horror video scene recognition algorithm is necessary for web filtering.

### 1.1 Horror video scene recognition review

Horror movies, a major component of horror material in the Web, are films that strive to elicit the emotions of fear, horror and terror from viewers. The earlier work on horror video scene recognition can be dated back to a part of affective video scene classification [2–4] whose final goal is to categorize movie scenes based on human emotions.

As an emerging problem, horror scene recognition attracts more researchers' special attention [5–8] with its own characteristics. Wang et al. [5] firstly use Support Vector Machines (SVM) to identify horror scene based on several effective holistic features inspired by emotional perception theory. But they further find that the holistic features inevitably weaken the features of the real horror frames because the horror scenes sometimes contain several rather than most horror frames. To avoid this confusion, both Wang et al. [6] and Wu et al. [7] introduce the multi-instance learning (MIL) into horror scene recognition, in which the scene is represented as a bag of independent frames.

Either the holistic methods or MIL based methods only focus on independent frames without considering the underlying contextual cues in the video scene. However, as Li et al. [8] point out, the horror emotion recognition can benefit from the proper use of contextual cues. The independent frame cues and contextual cues among frames can be treated as different views of a horror video scene. For example, scenes that express horror emotion through gory shot can be recognized mostly depending on their independent frame cues, while scenes that express their horror affection through scenario need more contexts for recognition. So, an effective MIL algorithm for horror video scene recognition should consider both of these two types simultaneously.

### 1.2 Multi-instance learning Review

As a variant of supervised learning framework, Multiple Instance Learning (MIL) represents a sample with a bag of several instances instead of a single instance. It only gives each bag, not each instance, a discrete or real-value label. In binary classification case, the bag is considered to be positive if at least one instance is positive, and is considered to be negative if all instances are negative.

Past decades have witnessed great progress in mathematical models for the MIL problem, from axis-parallel concepts [9] to Diverse Density method [10], k-Nearest Neighbor based algorithm Citation-kNN [11], Expectation-Maximization version of Diverse Density(EMDD) [12], and MI-kernel method [13]. In the MI-kernel algorithm, Gartner et al. regard each bag as a set of feature vectors and then apply set kernel directly. Andrews et al. [14] proposed mi-SVM and MI-SVM through extending Support Vector Machine (SVM). However, as Zhou and X-u [15] indicated, all these MIL algorithms always treated the instances in a bag as independently and identically distributed (i.i.d), which impairs the performance of classification. Therefore, Zhou et al. [16] proposed two multi-instance learning methods, miGraph and MIGraph, which treat the instances non-i.i.d through defining the structure information with $\epsilon$-graph. Although both i.i.d MIL and non-i.i.d MIL methods have been proposed, there is no MIL model that integrates the independent and contextual information in a bag together.

### 1.3 Our work

In order to simultaneously consider both context and independent instance in a bag, we propose a Multi-view Multi-instance learning model (M$^2$IL) through sparse coding technique. The ’view’ in this paper means ’observing’ the same object (bag in MIL) from different viewpoints (contextual or independent). The framework of the proposed method is shown in Fig.1. In independent view, the bag is treated as a bag of independent instances without any interplay. The contextual view takes the bag as a structural pattern via $\epsilon$-graph by integrating the context cues among instances. These two patterns under different view can be mapped into different feature spaces using different kernels. Finally, they are integrated into a unified learning framework based on joint sparse coding for multi-instance classification, especially in horror video scene recognition.

The remainder of this paper is organized as follows. We briefly introduce the sparse coding technique in section 2. Section 3 gives out the details of proposed Multi-view Multi-instance leaning (M$^2$IL) model. Horror Video Scene Recognition based on M$^2$IL is presented in section 4. The experimental results and analysis are reported in section 5. Section 6 concludes this paper.

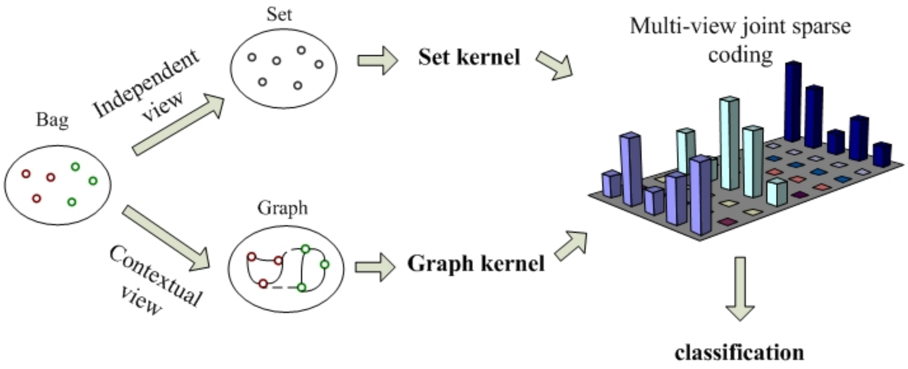

**Fig. 1.** The framework of proposed method.

## 2 Sparse Coding Review

Because sparse coding is the basis of the proposed algorithm, we start with a brief overview of it. The goal of sparse coding is to sparsely represent input vectors approximately as a weighted linear combination of a number of 'basis vectors'. Concretely, given input vector $x \in R^k$ and basis vectors $U = [u_1, u_2, \cdots, u_n] \in R^{k \times n}$, the goal of sparse coding is to find a sparse vector of coefficients $\alpha \in R^n$, such that $x \approx \mathbf{U}\alpha = \Sigma_j u_j \alpha_j$. It equals to solving the following objective:

$$\min_{\alpha} \|x - \mathbf{U}\alpha\|^2 + \lambda\|\alpha\|_1, \qquad (1)$$

where the first term of Eq.(1) is the reconstruction error, and the second term is used to control the sparsity of the coefficients vector $\alpha$. Regularization coefficient $\lambda$ controls the sparsity of $\alpha$. The larger $\lambda$ implies the sparser solution of $\alpha$.

## 3 Multi-view Multi-instance Learning (M$^2$IL) via Sparse Coding

First, we introduce the sparse coding based MIL from independent and contextual views respectively. Then, the details of the M$^2$IL model are represented.

### 3.1 MIL from independent view via sparse coding

According to definition of MIL, given a training data set $\{(X_1, y_1), \cdots, (X_i, y_i), \cdots, (X_N, y_N)\}$, where $X_i = \{x_{i1}, \cdots, x_{ij}, \cdots, x_{in_i}\} \subseteq \chi$ is called a bag and $y_i \in Y = \{-1, +1\}$ is the label of $X_i$, the $x_{ij} \in \chi$ is an instance, $N$ is the number of training bags, $n_i$ is the number of instances in $X_i$. If there exists $g \in \{1, \cdots, n_i\}$ such that $x_{ig}$ is a positive instance, then $X_i$ is a positive bag and thus $y_i = +1$,otherwise $y_i = -1$.

From independent viewpoint, a bag in MIL is treated as a loose 'set' that includes independent instances. A test bag is given as $X_{test} = \{x_{test1}, \cdots, x_{test2}, \cdots, x_{testn_{test}}\} \subseteq \chi$. Borrowing the sparse coding technique, we can sparsely linearly reconstruct the test bag using the training bags. Unfortunately, the test bag cannot directly be represent by the training bags based on sparse coding as Eq.(1) due to its set structure. So, we apply a feature mapping function $\varphi : X \to R^d$ to map the set patter $X$ to a high dimensional feature space as: $X \to \varphi(X)$. Thus we obtain a basis matrix $\mathbf{B} = [\varphi(X_1), \varphi(X_2), \cdots \varphi(X_N)]$. Then we define the sparse coding in high dimensional feature space as:

$$\min_{\alpha} \|\varphi(X_{test}) - \mathbf{B}\alpha\|^2 + \lambda' \|\alpha\|_1 \qquad (2)$$

### 3.2 MIL from contextual view via sparse coding

In this section, we address the bag in MIL from contextual viewpoint via sparse coding, in which graph is introduced to model the instances and their contextual relationship. Inspired by [16], we build a $\epsilon$-graph, which is shown to be helpful [17] in discovering the underlying manifold structure of data, to model the context among instances in each bag. For a bag $X_i$, $\mathbf{W}_i \in R^{n_i \times n_i}$ is set as a $\epsilon$-graph adjacency weight matrix. We compute the distance of every pair of instance nodes, e.g. $x_{i,k}$ and $x_{i,l}$. If the distance between $x_{i,k}$ and $x_{i,l}$ is smaller than a pre-set threshold $\varepsilon$, then an edge is established between these two nodes, and the weight value $\mathbf{W}^i_{kl}$ in $\mathbf{W}^i$ is set as 1, otherwise 0. Now, a bag of instances of $X_i$ are reconstructed as a $\epsilon$-graph $G_i$.

Then, the training data is represented as $\{(X_1, G_1, y_1), \cdots, (X_i, G_i, y_i), \cdots, (X_N, G_N, y_N)\}$, and a test bag is also given as $(X_{test}, G_{test}, y_{test})$. Similarly, the test graph cannot directly be represented by the training bags based on sparse coding as Eq.(1). We apply another feature mapping function $\phi : G \to R^d$ to map the graph pattern $G$ to a high dimensional feature space as: $G \to \phi(G)$. Thus the basis matrix $\mathbf{U}$ in Eq.(1) can be replace by $\mathbf{C} = [\phi(G_1), \phi(G_2), \cdots, \phi(G_n)]$.

And the sparse coding in Eq.(1) can be rewritten in high dimensional feature space as :

$$\min_{\beta} \|\phi(G_{test}) - \mathbf{C}\beta\|^2 + \lambda" \|\beta\|_1 \tag{3}$$

### 3.3 Multi-view Multi-instance Learning via joint sparse coding

**A. Multi-view Multi-instance Learning via sparse coding.** As discussed in section 1, both independent and contextual views are necessary in most MIL, especially for horror video scene recognition. Therefore, we propose a Multi-view Multi-instance Learning model by integrating them into a unified joint sparse coding framework based on the $\ell_{2,1}$ norm:

$$\min_{\alpha',\beta'} \frac{1}{2}\|\varphi(X_{test}) - \mathbf{B}\alpha'\|^2 + \frac{1}{2}\|\phi(G_{test}) - \mathbf{C}\beta'\|^2 + \eta\|[\alpha' \quad \beta']\|_{2,1} \tag{4}$$

In this model the first two terms are the reconstruction error from independent and contextual views respectively, and the third term is regularization term to control coefficients' sparsity.

**B. Optimization for M$^2$IL model.** In this part, we discuss how to optimize the object function in Eq.(4). First, we make some reformulation on Eq.(4). Suppose that we have a training set with $M$ classes, we can group them into training matrices $\mathbf{B}$ and $\mathbf{C}$ according to class labels, $\mathbf{B} = [\mathbf{B}_1 \cdots \mathbf{B}_m \cdots \mathbf{B}_M]$ and $\mathbf{C} = [\mathbf{C}_1 \cdots \mathbf{C}_m \cdots \mathbf{C}_M]$. Accordingly, we group $\alpha'$ and $\beta'$ as $[{\alpha'_1}^T \cdots {\alpha'_m}^T \cdots {\alpha'_M}^T]^T$ and $[{\beta'_1}^T \cdots {\beta'_m}^T \cdots {\beta'_M}^T]^T$. Let $q_m = [\alpha'_m \quad \beta'_m]$, $q^1 = \alpha'$, $q^2 = \beta'$, $\mathbf{Q} = [q^r_m]_{m,r}$, $m = 1, \cdots, M$ $r = 1, 2$, then Eq.(4) can be rewritten as:

$$\min_{Q} \frac{1}{2}\|\varphi(X_{test}) - \Sigma_{m=1}^{M}\mathbf{B}_m q^1_m\|^2 + \frac{1}{2}\|\phi(G_{test}) - \Sigma_{m=1}^{M}\mathbf{C}_m q^2_m\|^2 + \eta \sum_{m=1}^{M} \|q_m\|_2 \tag{5}$$

Then, the $\ell_{2,1}$ mixed-norm Accelerated Proximal Gradient (APG) algorithm proposed by [18] is introduced. The algorithm alternately updates a weight matrix sequence $\{\hat{\mathbf{Q}}^t = [q^{r,t}_m]\}_{t\geq 1}$ and an aggregation matrix sequence $\{\hat{\mathbf{V}}^t = [v^{r,t}_m]\}_{t\geq 1}$. Each iteration consists of two steps as:

**1)A generalized gradient mapping step.** Given the current matrix $\hat{\mathbf{V}}^t$, $\hat{\mathbf{Q}}^{t+1}$ is updated as follows:

$$\begin{aligned} \hat{q}^{r,t+1} &= \hat{v}^{r,t} - \mu\nabla^{r,t}, \quad r = 1, 2 \\ \hat{q}^{t+1}_m &= [1 - \frac{\eta\mu}{\|\hat{q}^{t+1}_m\|_2}]_+ \hat{q}^{t+1}_m, \quad m = 1, \cdots, M \end{aligned} \tag{6}$$

where

$$\nabla^{1,t} = -\mathbf{B}^T\varphi(X_{test}) + \mathbf{B}^T\mathbf{B}\hat{v}^{1,t} \tag{7}$$

$$\nabla^{2,t} = -\mathbf{C}^T\phi(G_{test}) + \mathbf{C}^T\mathbf{C}\hat{v}^{2,t} \tag{8}$$

$\mu$ is the step size parameter, and $[\cdot]_+ = max(\cdot, 0)$.

**2) The aggregation step.** $\hat{\mathbf{V}}^{t+1}$ is updated by constructing a linear combination of $\hat{\mathbf{Q}}^t$ and $\hat{\mathbf{Q}}^{t+1}$ as follows:

$$\hat{\mathbf{V}}^{t+1} = \hat{\mathbf{Q}}^{t+1} + \frac{\tau_{t+1}(1-\tau_t)}{\tau_t}(\hat{\mathbf{Q}}^{t+1} - \hat{\mathbf{Q}}^t) \tag{9}$$

The sequence $\{\tau_t\}_{t\geq 1}$ is conventionally set as $\tau_t = 2/(t+2)$ [19].

In the optimization algorithm, the key step is to compute $\nabla^{r,t}(r = 1, 2)$ according to Eq.(7)and(8) in which the matrix $\mathbf{B} = [\varphi(X_1), \varphi(X_2), \cdots, \varphi(X_N)]$ and $\mathbf{C} = [\phi(G_1), \phi(G_2), \cdots, \phi(G_N)]$ are the training features mapped from the original space to the high dimensional space. The $\mathbf{B}^T\mathbf{B}$, $\mathbf{B}^T\varphi(X_{test})$ in Eq.(7) and $\mathbf{C}^T\mathbf{C}$, $\mathbf{C}^T\phi(G_{test})$ in Eq.(8) can be represented as:

$$\begin{aligned}
\mathbf{B}^T\mathbf{B} &= \begin{bmatrix} \varphi(X_1)^T\varphi(X_1) & \varphi(X_1)^T\varphi(X_2) & \cdots & \varphi(X_1)^T\varphi(X_N) \\ \varphi(X_2)^T\varphi(X_1) & \varphi(X_2)^T\varphi(X_2) & \cdots & \varphi(X_2)^T\varphi(X_N) \\ \cdots & & & \cdots \\ \varphi(X_N)^T\varphi(X_1) & \varphi(X_N)^T\varphi(X_2) & \cdots & \varphi(X_N)^T\varphi(X_N) \end{bmatrix} = \mathbf{K}^1, \\
\mathbf{B}^T\varphi(X_{test}) &= \begin{bmatrix} \varphi(X_1)^T\varphi(X_{test}) \\ \varphi(X_2)^T\varphi(X_{test}) \\ \cdots \\ \varphi(X_N)^T\varphi(X_{test}) \end{bmatrix} = \mathbf{H}^1
\end{aligned} \tag{10}$$

$$\begin{aligned}
\mathbf{C}^T\mathbf{C} &= \begin{bmatrix} \phi(G_1)^T\phi(G_1) & \phi(G_1)^T\phi(G_2) & \cdots & \phi(G_1)^T\phi(G_N) \\ \phi(G_2)^T\phi(G_1) & \phi(G_2)^T\phi(G_2) & \cdots & \phi(G_2)^T\phi(G_N) \\ \cdots & & & \cdots \\ \phi(G_N)^T\phi(G_1) & \phi(G_N)^T\phi(G_2) & \cdots & \phi(G_N)^T\phi(G_N) \end{bmatrix} = \mathbf{K}^2, \\
\mathbf{C}^T\phi(G_{test}) &= \begin{bmatrix} \phi(G_1)^T\phi(G_{test}) \\ \phi(G_2)^T\phi(G_{test}) \\ \cdots \\ \phi(G_N)^T\phi(G_{test}) \end{bmatrix} = \mathbf{H}^2
\end{aligned} \tag{11}$$

From Eq.(10) and Eq.(11), we find that $\mathbf{B}^T\mathbf{B}$ and $\mathbf{C}^T\mathbf{C}$ are the dot products in high dimensional feature space between the training bags; while $\mathbf{B}^T\varphi(X_{test})$ and $\mathbf{C}^T\phi(G_{test})$ are dot products between the training bags and the test bags. Consequently, kernel matrixes $\mathbf{K}^1$, $\mathbf{K}^2$, $\mathbf{H}^1$ and $\mathbf{H}^2$ can be used to represent $\mathbf{B}^T\mathbf{B}$, $\mathbf{C}^T\mathbf{C}$, $\mathbf{B}^T\varphi(X_{test})$ and $\mathbf{C}^T\phi(G_{test})$, respectively. So the kernel function definition is important for computations of kernel matrixes $\mathbf{K}^1$, $\mathbf{K}^2$, $\mathbf{H}^1$ and $\mathbf{H}^2$. We respectively define the kernel function according to [13] [16] as follows:

$$K^1(X_i, X_j) = \varphi(X_i)^T\varphi(X_j) = \frac{\sum\limits_{a=1}^{n_i}\sum\limits_{b=1}^{n_j} Ker(x_{i,a}, x_{j,b})}{\sum\limits_{l=1}^{n_i}\sum\limits_{l'=1}^{n_i} Ker(x_{i,l}, x_{i,l'})\sum\limits_{s=1}^{n_j}\sum\limits_{s'=1}^{n_j} Ker(x_{j,s}, x_{j,s'})} \tag{12}$$

$$K^2(G_i, G_j) = \phi(G_i)^T\phi(G_j) = \frac{\sum\limits_{a=1}^{n_i}\sum\limits_{b=1}^{n_j} \omega_{i,a}\omega_{j,b}Ker(x_{i,a}, x_{j,b})}{\sum\limits_{a=1}^{n_i}\omega_{i,a}\sum\limits_{b=1}^{n_j}\omega_{j,b}} \tag{13}$$

where

$$Ker(x_{i,a}, x_{j,b}) = \exp(-\sigma\left\|x_{i,a} - x_{j,b}\right\|^2) \tag{14}$$

is the Gaussian radial basis function(RBF) kernel. And $\omega_{i,a} = 1/\sum_{u=1}^{n_i}\mathbf{W}^i_{a,u}$, $\omega_{j,b} = 1/\sum_{u=1}^{n_j}\mathbf{W}^j_{b,u}$, $\mathbf{W}^i$ and $\mathbf{W}^j$ are the adjacency weights matrixes for bag $X_i$ and $X_j$, respectively. Then, the $\nabla^{r,t}$ is computed as $\nabla^{r,t} = -\mathbf{H}^r + \mathbf{K}^r\hat{v}^{r,t}$,$r = 1, 2$.

**C. Bag classification.** After getting the coefficients matrix $\mathbf{Q}$, the reconstruction residual $r_m(X_{test})$ of the test bag in class $m \in 1, \cdots, M$ is defined as:

$$\begin{aligned} r_m(X_{test}) &= \left\|\varphi(X_{test}) - \mathbf{B}_m q^1_m\right\|_2^2 + \left\|\phi(G_{test}) - \mathbf{C}_m q^2_m\right\|_2^2 \\ &= \sum_{r=1}^{2}((\delta_m(q^r))^T\mathbf{K}^r\delta_m(q^r) - 2\mathbf{H}^r\delta_m(q^r)) \end{aligned} \tag{15}$$

$$[\delta_m(q^r)]_l = \begin{cases} (q^r)_l & y_l = m \\ 0 & y_l \neq m \end{cases}$$

where $\delta_m(q^r)$ is a coefficients selector that only selects coefficients associated with class $m$. The final class $c$ that is assigned to the test bag $X_{test}$ is the one that gives the smallest residual, as:

$$c = \arg\min_m(r_m(X_{test})) \tag{16}$$

## 4 Horror Video Scene Recognition based on M²IL

In this section, we detail horror video scene recognition based on M$^2$IL.

### 4.1 Bag construction

Given $N$ video scenes, each scene $VS_i$ is first divided into $n_i$ shots as $S_{i,1}, S_{i,2}$, $\cdots, S_{i,n_i}$ through measuring information transported from one frame to another by Mutual Information(MI) [20]. Then key frames are extracted as $F_{i,1}, F_{i,2}, \cdots$,

$F_{i,n_i}$ each of which is the central frame of every shot, and the visual and audio features $f_{i,j} \in R_m$ of each frame $F_{i,j}$ are extracted. Now the 'video scene bag' is constructed from these key frames and their features, shown in Figure 2.

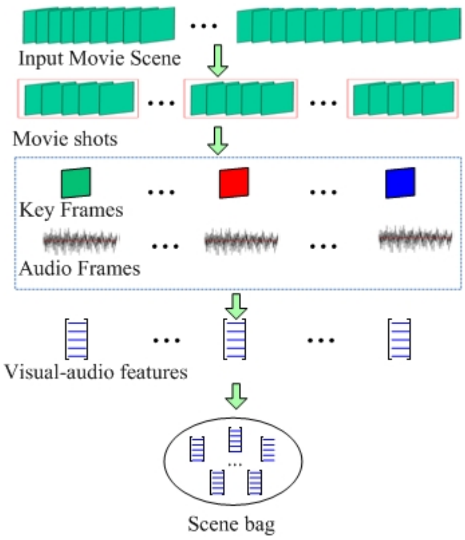

**Fig. 2.** Bag construction of each video scene.

### 4.2 Feature extraction

The features extracted from video scene are based on the emotional perception theory. In order to compare with the methods in [6], the same visual and audio features are used in this paper. They are summarized in Table 1.

### 4.3 Recognition

Given a group of the training set, the label for each video scene in the training set is set as 1 if it is a horror scene, set as -1 otherwise. The M$^2$IL is applied to identify the test horror scene from both independent and contextual viewpoints.

## 5 Experiments

The experiments in this paper include two parts: the first one focues on horror scene recognition, the second one is conducted on general MIL data sets to validate the effect of the proposed M$^2$IL model.

**Table 1.** Summary of all features.

| Feature Type | Features | Shot description |
|---|---|---|
| Audio | Mel-Frequency Cepstral Coefficients (MFCCs) | Illustrates the mel-cepstral features and is computed from the Fast Fourier Transform(FFT) power coefficients. |
| | Spectral power | Used to measure the energy intensity of audio signal. |
| | Spectral centroid | A measurement of music brightness. |
| Visual | Emotional intensity | Color emotion models developed by Ou et al. [21]. |
| | Color harmony | A quantitative two-color harmony model developed by Ou et al. [22]. |
| | Variance of color | Determinant of covariance matrix of $L$, $u$, $v$ color space of each key frame |
| | Lighting key | Median of $L$ value of the Luv color space and mean of proportion of pixels whose lightness are below a certain shadow threshold. |
| | Texture | Based on a six-stimulus basis for stochastic texture perception [23]. |

### 5.1 Horror scene data set

We download a large number of movies from the internet. These movies 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 video scenes and 400 non-horror video scenes in total. The proposed method is compared with MIL based horror video scene recognition method proposed by Wang et al. [6], miGraph method proposed by Zhou et al. [16] and MI-kernel method proposed by Gartner et al. [13]. The average accuracies of ten times 10-fold cross validation is used as the final performances for each method.

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.(17) are used to evaluate the performances.

$$P = \frac{|HS \cap ES|}{|ES|}, R = \frac{|HS \cap ES|}{|HS|}, F_1 = \frac{2 \times P \times R}{P + R} \tag{17}$$

The average Precision ($P$), Recall($R$) and F-measure ($F_1$) are shown in Table 2. The methods MI-SVM, CKNN, EM-DD, SI-SVM in Table 2 denote the MIL based recognition methods with different MIL classifiers [6], and the results are cited from [6] because almost same dataset is used in this paper.

The results in Table 2 show that the performances of M$^2$IL and miGraph methods outperform other MIL based methods since they consider context cues inside a scene. It shows that the context interplay is very useful in horror scene

**Table 2.** Experiment results on horror scene data(%)

| **Algorithm** | **Precision**($P$) | **Recall**($R$) | **F-measure**($F_1$) |
|---|---|---|---|
| M$^2$IL | 85.47±0.49 | 85.19±0.38 | 85.33±0.33 |
| miGraph | 81.87±1.95 | 82.4±1.25 | 82.14±1.2 |
| MI-kernel | 80.7±1.42 | 81.43±0.9 | 81.05±0.5 |
| 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 |

recognition. On the other hand, the fact that performance of M$^2$IL is much better than miGraph and MI-kernel shows that horror scene recognition can benefit from considering multi-views rather than only one view. In addition, the lower standard deviations of M$^2$IL implies its stableness. Furthermore, the training free character embedded in the sparse coding classifier makes it possible to extend M$^2$IL as an online classifier that is necessary for many video analysis applications.

### 5.2 Experiment on MIL data sets

To evaluate recognition performance of our proposed M$^2$IL method, we also apply it on the general MIL data sets. Two popular MIL data sets are adopted in this paper. The first data set includes five benchmark data sets that are widely used in the studies of multi-instance learning, including Musk1, Musk2, Elephant, Fox and Tiger. Musk1 contains 47 positive and 45 negative bags, Musk2 contains 39 positive and 63 negative bags, and each of the other three data sets contains 100 positive and 100 negative bags. More details of these five data sets can be found in [9, 14]. The second set is an image categorization set. It includes two subsets: 1000-Image set and 2000-Image set that contain ten and twenty categories of COREL images, respectively. Each category of these two image subsets has 100 images. Each image is regarded as a bag, and the ROIs (Region of Interests) in the image are regarded as instances described by nine features [24, 25].

**A. Results on Benchmark data sets.** In this section, we compare M$^2$IL with miGraph, MIGraph and MI-Kernel via repeating 10-fold cross validations ten times through following the same procedure described in [16]. The average test accuracy and standard deviations are shown in Table 3. The experimental results of other methods, including MI-SVM and mi-SVM [14], MissSVM [15], PPMM kernel [26], the Diverse Density algorithm [10] and EM-DD [12], are cited from the work of Zhou et al. [16].

Table 3 shows that the performance of M$^2$IL is pretty good. It achieves better performances than MIGraph and miGraph on Musk1, Elephant and Fox

**Table 3.** Accuracy (%) on benchmark sets

| **Algorithm** | ***Musk1*** | ***Musk2*** | ***Elephant*** | ***Fox*** | ***Tiger*** |
|---|---|---|---|---|---|
| M$^2$IL | 91.6±2.8 | 90.6±1.3 | 88.5±1.1 | 62.7±1.8 | 86.8±1.2 |
| miGraph | 88.9±3.3 | 90.3±2.6 | 86.8±0.7 | 61.6±2.8 | 86.0±1.6 |
| MIGraph | 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 |
| mi-SVM | 87.4 | 83.6 | 82 | 58.2 | 78.9 |
| missSVM | 87.6 | 80.0 | N/A | N/A | N/A |
| PPMM | 95.6 | 81.2 | 82.4 | 60.3 | 82.4 |
| DD | 88.0 | 84.0 | N/A | N/A | N/A |
| EMDD | 84.8 | 84.9 | 78.3 | 56.1 | 72.1 |

**Table 4.** Accuracy (%) on Image Categorization

| **Algorithm** | ***1000-Image*** | ***2000-Image*** |
|---|---|---|
| M$^2$IL | 84.3:[83.2,84.8] | 67.1:[66.8,67.3] |
| miGraph | 82.4:[80.2,82.6] | 70.5:[68.7,72.3] |
| MIGraph | 83.9:[81.2,85.7] | 72.1:[71.0,73.2] |
| MI-kernel | 81.8:[80.1,83.6] | 72.0:[71.2,72.8] |
| MI-SVM | 74.7:[74.1,75.3] | 54.6:[53.1,56.1] |
| DD-SVM | 81.5:[78.5,84.5] | 67.5:[66.1,68.9] |
| missSVM | 78.0:[75.8,80.2] | 65.2:[62.0,68.3] |
| Kmeans-SVM | 69.8:[67.9,71.7] | 52.3:[51.6,52.9] |
| MILES | 82.6:[81.4,83.7] | 68.7:[67.3,70.1] |

sets. The performances of M$^2$IL, MIGraph, miGraph and MI-kernel on Musk2 and Tiger are comparable. In addition, we can notice that the proposed M$^2$IL has lower standard deviations on different benchmark sets, which indicates the stableness of M$^2$IL.

**B. Results on Image Categorization.** The second experiment is conducted on the two image categorization sets. We use the same experimental routine as that described in [24]. For each data set, we randomly partition the images within each category in half, and use one subset for training and leave the other one for testing. The experiment is repeated five times with five random splits, and the average results are recorded. The overall accuracy as well as 95% confidence intervals is also provided in Table 4. For reference, the table also shows the best results of some other MIL methods that are given out by Zhou et al. [16].

From table 4, we can find that the M$^2$IL has comparable performances to miGraph on 1000-Image and 2000-Image sets, which again validates the effect of our method. Although the proposed M$^2$IL has better performances than most MIL methods without considering different views, the accuracy of M$^2$IL is slightly lower than miGraph and MI-kernel on 2000-Image sets. By analyzing and comparing the results in table 3 and table 4, we may obtain an observation that the M$^2$IL has relatively lower performances than miGraph and MI-Kernel when facing classification with too many categories. However, as a good alternative MIL method, it is seen that it has well performance in horror scene recognition.

## 6 Conclusion

Most existing MIL studies on horror scene recognition neglect the fact that there are multi-views in video data including contextual view and independent instance view. Integrating these two views effectively plays an important role in image emotion recognition. In this paper, we have proposed a novel Multi-view Multi-instance learning (M$^2$IL) model based on sparse coding to fuse different views into a unified framework. The experiments on both horror video dataset and general MIL dataset have shown that our model is not only superior to other existing horror recognition methods but also effective in other general Multi-instance problem.

## References

1. Hu, W.M., Wu, O., Chen, Z.: Recognition of pornographic web pages by classifying texts and images. IEEE TPAMI **29** (2007) 1019–1034
2. Hanjalic, A., Xu, L.Q.: Affective video content representation and modeling. IEEE TM **7** (2005) 143–154
3. Wang, H.L., Cheong, L.: Affective understanding in film. IEEE TCSVT **16** (2006) 689–704
4. Kang, H.B.: Affective content detection using hmms. ACM MM (2003) 259–262
5. Wang, J.C., Li, B., Hu, W.M., et al.: Horror movie scene recognition based on emotional perception. ICIP (2010) 1489–1492
6. Wang, J.C., Li, B., Hu, W.M., et al.: Horror video scene recognition via mutiple-instance learning. ICASSP (2011) 1325–1328
7. Wu, B., Jiang, X., Sun, T., et al.: A novel horror scene detection scheme on revised multiple instance learning model. MMM (2011) 377–388
8. Li, B., Xiong, W.H., Hu, W.M.: Web horror image recognition based on context-aware multi-instance learning. ICDM (2011) 1158–1163
9. Dietterich, T.G., Lathrop, R.H., Lozano-Perez, T.: Solving the multiple-instance problem with axis-parallel rectangles. Artif. Intell. **89** (1997) 1158–1163
10. Maron, O., Lozano-P'erez, T.: A framework for multiple-instance learning. Neural Inf. Process. Syst. **10** (1998) 570–576
11. Wang, J., Zucker, J.D.: Solving the multi-instance problem: A lazy learning approach. Proc. 17th Intl. Conf. Mach. Learn. (2000) 1119–1125
12. Zhang, Q., Goldman, S.A.: Em-dd: An improved multi-instance learning technique. Adv. Neural Inf. Process. Syst. **14** (2002) 1037–1080
13. Gartner, T., Flach, P.A., A.Kowalczyk, Smola, A.J.: Multi-instance kernels. Proc. 19th Intl. Conf. Mach. Learn. (2002) 179–186
14. Andrews, S., Tsochantaridis, I., Hofmann: Support vector machines for multiple-instance learning. Adv. Neural Inf. Process. **15** (2003) 561–568

15. Zhou, Z.H., Xu, J.M.: On the relation between multi-instance learning and semi-supervised learning. Proc. 24th Intl. Conf. Mach. Learn. (2007) 1167–1174
16. Zhou, Z., Sun, Y., Li, Y.: Multi-instance learning by treating instances as non-i.i.d. samples. ICML (2009) 1249–1256
17. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction. Science **290** (2000) 2319–2323
18. Yuan, X., Yan., S.: Visual classification with multi-task joint sparse representation. CVPR2010 (2010) 3493–3500
19. Tseng, P.: On accelerated proximal gradient methods for convex-concave optimization. submitted to SIAM Journal of Optimization (2008)
20. Cernekova, Z., Pitas, I., Nikou, C.: Information theory-based shot cut/fade detection and video summarization. IEEE TCSVT **16** (2006) 82–91
21. Ou, L., Luo, M., Woodcock, A., Wright, A.: A study of colour emotion and colour preference.part i: Colour emotions for single colours. Color Research  Application **29** (2004) 232–240
22. Ou, L., Luo, M.: A colour harmony model for two-colour combinations. Color Research  Application **31** (2006) 191–204
23. Geusebroek, J., Smeulders, A.: A six-stimulus theory for stochastic texture. International Journal of Computer Vision **62** (2005) 7–16
24. Chen, Y., Bi, J., Wang, J.Z.: Miles: Multiple-instance learning via embedded instance selection. IEEE TPAMI **28** (2006) 1931–1947
25. Chen, Y., Wang, J.Z.: Image categorization by learning and reasoning with regions. J. Mach. Learn. Res. **5** (2004) 913–939
26. Wang, H.Y., Yang, Q., Zhang, H.: Adaptive p-posterior mixturemodel kernels for multiple instance learning. ICML (2008) 1136–1143