Web Objectionable Video Recognition Based on Deep Multi-Instance Learning with Representative Prototypes Selection

Xinmiao Ding, Bing Li, Yangxi Li, Wen Guo, Yao Liu, Weihua Xiong, Weiming Hu

***Abstract*—To protect underage people from accessing objectionable videos in the Internet, an effective objectionable video recognition algorithm is necessary for web filtering. Recently, the multi-instance learning has been introduced for objectionable video recognition and achieves impressive results. However, hand-crafted features as well as redundant and noisy frames in objectionable videos become an intractable problem that inevitably degrades the recognition performance. In this paper, we propose a novel representative prototype selection algorithm embedding deep multi-instance representation learning. In the proposed method, an improved convolutional neural network is designed for multimodal multi-instance feature learning and a self-expressive dictionary learning model based on sparse and low rank constraint is designed to select the representative prototypes from each subspace of instances. Then the bag-level feature is constructed via mapping the bag to the selected prototypes. Experiments on three objectionable video sets show the effectiveness of our method for objectionable video recognition.**

***Index Terms*—Representative prototype selection, Objectionable video recognition, Deep learning.**

I. Introduction

With an exceptional boosting in the creation and propagation of multimedia content through Internet, the lack of control has allowed the distribution of many harmful documents related with pornography, violence, horror, racism, etc. To prevent people, especially sensitive social groups (e.g. children), from all kinds of harmful materials on the Internet, many content-based web filtering systems have been developed [1, 2, 3]. Effective recognition of objectionable videos is necessary for such web content security [4]. Recognition of objectionable videos is a newly emergent research topic in the context of excellent multimedia applications including multimedia retrieval [5, 6, 7, 8], multimedia content understanding [9, 10, 11, 12, 13], and multimodal fusion [14, 15, 16, 17], etc. In this paper, we focus on horror and violent video recognition which are still being under exploration.

A. Related Work

The definition of violence and horror is very subjective. In this paper, we simplify the concept by such definitions: violent scene will stimulate emotion impulses through showing the use of force to injure others or oneself, it usually includes fights, gun shots, explosions, and self-mutilation; while horror scene will strive to elicit the primary emotions of fear, horror, and terror, it usually includes serial killings, ghosts, monsters, vampires, animal killing, and irreligion [18]. The existing recognition methods for violent and horror videos can be classified into two types, video-based methods and frame-based methods, that differ in the way how the frames in a video are treated. Video-based method can be viewed as a global solution, while frame-based method can be viewed as a local one.

The video-based method treats a video as a single sample and extracts a global feature vector from the whole video without considering the content of a single frame in it. Yang et al. [19] presented a set of features for horror video recognition and studied the strength and disadvantage of classical affective recognition algorithms in horror video recognition. Inspired by

Manuscript received November 12, 2019; revised January 22, 2020; accepted May 09, 2016. This work is partly supported by the Beijing Natural Science Foundation(Grant No. JQ18018 and L172051), the Natural Science Foundation of China (Grant No. 61876100, U1803119, U1736106, 61906192, 61902401, 61572296, 61751212, 61721004 and 61772225), the NSFC-General Technology Collaborative Fund for Basic Research (Grant No. U1936204 and U1636218), the Science and Technology Service Network Initiative, CAS(Grant No. KFJ-STS-SCYD-317), the Key Research Program of Frontier Sciences, CAS (Grant No. QYZDJ-SSW- JSC040), the National Natural Science Foundation of Guangdong (Grant No. 2018B030311046) and Shandong Provincial Science Technology Support Program of Youth Innovation Team in Colleges(Grant No. 2019KJN041). The work of Bing Li was also supported by the Youth Innovation Promotion Association, CAS. (*Corresponding Author: Bing Li*)

Bing Li is with the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, also with PeopleAI, Inc., Beijing 100190, China, and also with the State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China (e-mail: bli@nlpr.ia.ac.cn).

Xinmiao Ding and Wen Guo are with the Shandong Technology and Business University, Yantai 264005, China (e-mail: dingxinmiao@126.com; wguo@ nlpr.ia.ac.cn).

Yangxi Li is with the National Computer Network Emergency Response Technical Team/Coordination Center of China (CNCERT/CC), Beijing 100029, China (e-mail: liyangxi@outlook.com).

Yao Liu is with the Beijing Institute of Applied Science and Technology, Beijing 100091, China (e-mail: liuyao88@mail.ustc.edu.cn).

Weihua Xiong is with the PeopleAI, Inc., Beijing 100190, China (e-mail: wallace.xiong@gmail.com).

Weiming Hu is with the CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Beijing 100190, China, also with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China, and also with the School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100039, China (e-mail: wmhu@nlpr. ia.ac.cn).

emotional perception theory, Wang et al. [20] extracted several effective holistic features and used support vector machine (SVM) to identify horror videos. Datta et al. [21] addressed the problem of detecting human violence in a video based on motion trajectory information and orientation information of a person's limbs. JeHo Nam et al. [22] exploited multiple "audio-visual" signatures to create a perceptual relation for conceptually meaningful violent video scene identification. Considering the importance of local spatio-temporal features in characterizing the multimedia content, De Souza [23] presented a violence detector built on the concept of visual codebooks, it used linear support vector machines that considered local spatio-temporal features with bags of visual words. To get discriminative features for the representation of violent video segments, Acar et al. [24] constructed mid-level audio features with vector quantization-based (VQ-based) method and sparse coding-based (SC-based) method. Based on Multiple Kernel Learning and the combination of visual and audio features, Shinichi et al. [25] introduced Mid-level Violence Clustering to solve violent scenes recognition. The big problem of the video-based method is that those global features extracted from a whole video are lack of discrimination because horror/violent scenes do not often show up in the majority frames.

To avoid such limitations, the frame-based methods follow the paradigm of multi-instance learning (MIL) which has been successfully used in visual applications and exploited new idea of traditional computer vision[26,27]. It treats each video as a bag of frames and labels it objectionable if there exists at least one objectionable frame. The features of each frame are extracted and considered by MIL algorithms during recognition. Such kind of solution can be dated back to the work of Wang et al. [28] and Wu et al. [29], who firstly introduced MIL into horror video recognition. In the MIL-based methods, a video is represented as a bag of key frames with corresponding independent features. To consider the multiple semantics of the objectionable video, Hu et al. [18] proposed Multi-Perspective Context-Aware Cost-Sensitive Multi-Instance Sparse Coding combining Multi-Instance Learning (MC-MI-J-SC) to recognize violent and horror videos. Hao et al. [30] classified the video into violent and non-violent using Multi-Modal Multi-Instance Learning and Attribute Discovery approach by combining audio-video with text information for web video recognition.

Although frame-based methods with MIL generally achieve better performance, it is still subject to two limitations. On the one hand, the features used in most existing methods are hand-crafted feature descriptors which usually require substantial prior domain knowledge. With the increasing popularity of the deep learning-based approaches, deep feature learning has shown much better performance in many visual applications [31,32,33,34]. There needs more advanced feature to capture high-level semantics in diverse unconstrained videos. On the other hand, the non-objectionable instances in objectionable bags bring a large amount of redundant and noise information that has no discrimination for recognition. How to effectively prune these useless instances and keep the discriminative and representative ones still remains a challenging problem for objectionable video recognition.

B. Our Method

To address the two problems of frame-based methods, we propose a novel representative prototype selection algorithm (MILRPS) embedding deep multi-instance representation learning and apply it to objectionable video recognition. In this method, a novel multi-instance convolutional neural network is proposed to learn discrimination features of each instance. Then a self-expressive dictionary learning model with low-rank and sparse constraints is proposed to select a set of representative prototypes. After bag feature mapping to these prototypes, a more effective objectionable video recognition system can be built.

The main contributions of our work are two-fold.

(i) The proposal of the deep multi-instance representation learning. Most existing deep representation learning methods are trained for single instance applications. They cannot be used directly in MIL applications due to the unknown labels of instances. To overcome this limitation, a novel CNN with new prediction loss is especially designed for MIL. It can be trained using bag's label instead of instances'. In addition, to fuse both visual and audio features, a two-channel structure is integrated to learn high level multimodal instance features.

(ii) The proposal of a novel instance selection framework embedding three constraints according to the characteristics of the selected representative prototypes:

- *Sparsity criteria* which ensures that the most informative instances will be selected.
- *Low rank criteria* which ensures that the selected prototypes will distribute among each subspace.
- *Error term **E*** which improves the robustness through modeling the noisy and possible corruption.

The remainder of this paper is organized as follows. We briefly introduce the overview of the proposed method in section II. Section III gives out the construction of deep multi-instance representation learning. The details of the prototype selection based on self-expressive dictionary learning are presented in section IV. The experimental results and analysis are reported in section V. Section VI concludes this paper.

II. System Overview

The objectionable video recognition proceeds in five main stages: shot segmentation and key frame extraction, deep multi-instance representation learning, representative prototypes selection, bag feature mapping and objectionable video recognition. Fig. 1 gives an overview of our framework.

Step 1: Shot segmentation and key frame extraction. Given a set of videos, we firstly divide each of them into shots through measuring mutual information (MI) transported from one frame to another [35]. Then central frame of every shot is extracted as key frame and audio of each shot is transformed into 2-D image based on the spectrogram [36].

Step 2: Deep multi-instance representation learning. The key frames and the spectrograms in a video are integrally fed into the designed multi-instance CNN (MI-CNN) to learn the deep instance feature. Two fully connected layers are designed for multimodal fusing and a new loss prediction based on the assumption of MIL is designed for back propagation. After training, we get the output of the second fully connected layer in the CNN as the high level feature of instance. Then, the video bag is considered as a set of deep instance features extracted from its key frames and audio spectrograms.

Step 3: Representative prototypes selection. To prune the redundancy and disturbance (non-objectionable instance) in bags, especially those positive bags (objectionable videos), we construct self-expressive dictionary learning to select representative prototypes. To make the representatives approach the distribution of instance subspaces to the utmost, we add sparse and low rank constraint to the self-expressive model. Meanwhile, an additional error term is also appended to improve robustness of proposed method when confronting data noise and corruption. We measure the importance of representatives through ranking their capacity and select top k representatives to compose the mapping prototype set (MPS).

Step 4: Bag feature mapping. Through bag feature mapping which computes the similarity between a bag and each mapping prototype in MPS, we can obtain the bag-level feature. Then a SVM classifier is trained on these bag-level features with corresponding bag labels.

Step 5: Objectionable video recognition. For a test video, after feature extraction and bag construction, we map the test video bag to MPS and get test bag's feature. Finally, we apply the trained SVM on a test video to estimate its category (objectionable or non-objectionable).

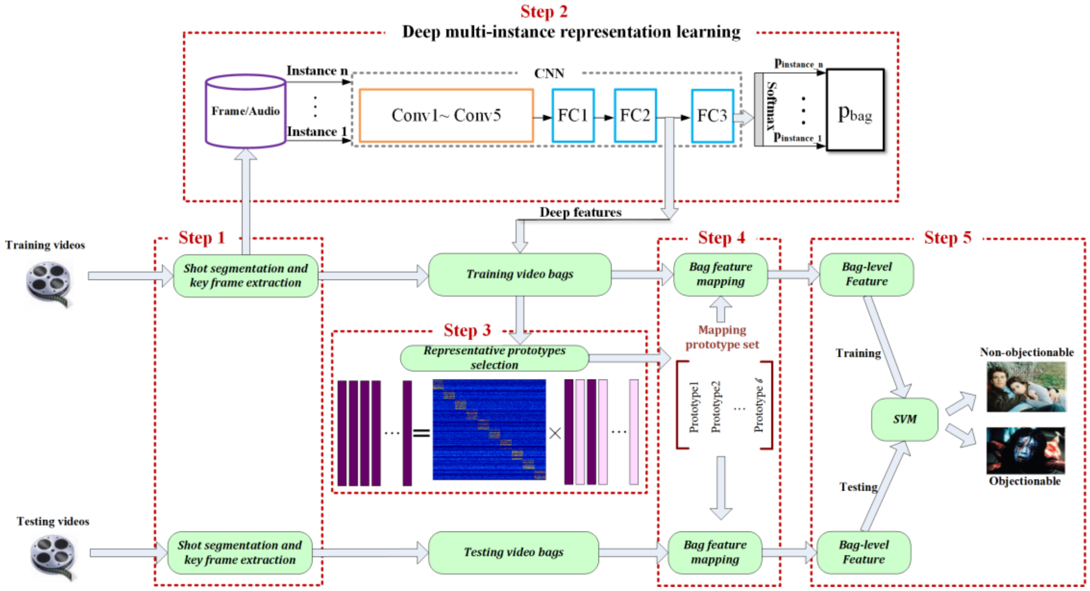

Fig.1. Framework of proposed method with five main stages. Step 1: shots in videos are segmented through measuring mutual information (MI) and key frames and audios are extracted from each shot; Step 2: key frames and audio spectrograms in a video are integrally fed into the multi-instance CNN (MI-CNN) to learning the deep instance feature; Step 3: self-expressive dictionary learning is constructed to select representative prototypes and build the mapping prototype set(MPS); Step 4: bag-level feature is obtained by computing the similarity between a bag and each mapping prototype in MPS; Step5: the SVM trained by mapping features is applied to a test video to estimate its category (objectionable or non-objectionable).

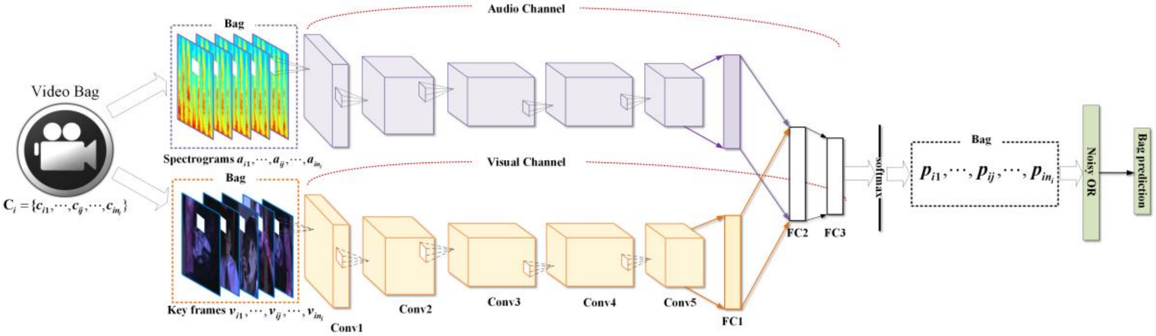
Audio Channel
Bag
Video Bag
Spectrograms $a_{i1},\cdots,a_{ij},\cdots,a_{in_i}$
$C_i=\{c_{i1},\cdots,c_{ij},\cdots,c_{in_i}\}$
Visual Channel
Bag
Key frames $v_{i1},\cdots,v_{ij},\cdots,v_{in_i}$
Conv1
Conv2
Conv3
Conv4
Conv5
FC1
FC2
FC3
softmax
Bag
$p_{i1},\cdots,p_{ij},\cdots,p_{in_i}$
Noisy OR
Bag prediction

Fig.2. Framework of MI-CNN. There are two channels to learn the visual and audio features respectively. Each channel is a CNN with five convolutional layers (Conv1~Conv5) followed by a pooling layer and three fully connected layers(FC1~FC3). To fuse the multimodal features, the visual and audio channels are combined at the last two fully connected layers (FC2 and FC3). For each channel, the audio or visual instances in a video are fed into the net successively and the label prediction of each instance is computed through softmax layer after FC3 in sequence. Finally, the label of bag is predicted by NoisyOR layer with all predicted labels of instances in it.

III. Deep Multi-instance representation learning

In this section, we present a novel deep multi-instance representation learning method. Before giving out the details, we briefly review the definition of MIL. Given a training data set $\{(\mathbf{C}_1, y_1), \cdots, (\mathbf{C}_2, y_i), \cdots, (\mathbf{C}_N, y_N)\}$, where $\mathbf{C}_i = \{c_{i1}, \cdots,$

$c_{ij}, \cdots, c_{in_i}\}$ is a video bag and $y_i \in Y = \{-1, +1\}$ is its label. The c_{ij} is an instance in bag $\mathbf{C}_i$ which specifically refers to the key frame as well as its contextual audio spectrogram. N is the number of training bags, n_i is the number of instances in $\mathbf{C}_i$. If there exists $g \in \{1, \cdots, n_i\}$ such that c_{ig} is a positive instance, then $\mathbf{C}_i$ is a positive bag and thus $y_i = +1$; otherwise $y_i = -1$.

A. Deep MIL Framework

Considering the success of the AlexNet [33] in visual applications, we use it as our basic architecture to learn the instance feature. The proposed deep MIL framework is as shown in the Fig. 2. There are two channels to learn the visual and audio features respectively. Each channel is a CNN with five convolutional layers followed by a pooling layer and three fully connected layers. To fuse the multimodal features, the visual and audio channels are combined at the last two fully connected layers (FC2 and FC3).

It is well known that the input of CNN is single instance and the output of the last fully connected layer with a softmax is the label prediction of the instance. Through optimization of the loss between prediction and target with back propagation, the weights of the network can be learned. However, the instance label in MIL is unknown so that the traditional training process cannot be used for MIL paradigms. In this paper, we consider to utilize the label of bag to modify the training of the CNN network.

Based on the assumption of MIL, the bag label is determined by the labels of its instance. Consequently, in the training process, a bag of instances instead of one instance are inputted into the network and an additional prediction layer "Noisy OR" is added after the last fully connected layer of instance as shown in the Fig.2. Through the additional layer, the bag label can be predicted based on the label predictions of all instances in it. We call this modified CNN multi-instance CNN (MI-CNN).

To train the MI-CNN, a new prediction loss is designed to realize the back propagation. Inputting one training bag $(\mathbf{C}_i, y_i)$ into the network, for each of its instances c_{ij} $(j = 1, \dots, n_i)$ we can get layer-wise features from the first convolutional layer (Conv1) to the output of the last fully connected layer (FC3). Supposing the FC3 output of the jth instance is $f_{ij} \in \mathbb{R}$, followed by a softmax layer, f_{ij} is transformed into a probability distribution $p_{ij} \in \mathbb{R}$ for objects of binary classification as following:

$$p_{ij} = \frac{1}{1 + \exp(-f_{ij})} \tag{1}$$

However p_{ij} cannot be used to compute the loss for the unknown label of instance. Under the definition of MIL, if there exists one positive instance in the bag, then the bag is positive; otherwise, the bag is negative. We introduce the NoisyOR generative model [38] and predict the probability distribution p_i of a bag as a "Noisy OR" of instance probability:

$$p_i = 1 - \prod\nolimits_j (1 - p_{ij}) \tag{2}$$

Next, cross entropy is used to measure the prediction loss of the network. Specifically, we have

$$\text{Loss} = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \tag{3}$$

Then the gradients of the deep convolutional neural network are calculated via back propagation:

$$\begin{aligned}\frac{\partial \text{Loss}}{\partial f_{ij}} &= \frac{\partial \text{Loss}}{\partial p_i} \cdot \frac{\partial p_i}{\partial p_{ij}} \cdot \frac{\partial p_{ij}}{\partial f_{ij}} \\ &= -\left(\frac{y_i}{p_i} - \frac{1 - y_i}{1 - p_i}\right) \cdot \left[\prod\nolimits_{k \neq j} (1 - p_{ik})\right] \cdot \left[p_{ij}(1 - p_{ij})\right] \\ &= \frac{(p_i - y_i) p_{ij}}{p_i}\end{aligned} \tag{4}$$

According to (4), we can iteratively update the weights of networks using the back propagation until getting ideal loss.

B. Deep Instance Feature Extraction

For an unknown video bag, the MI-CNN can be directly used to predict its label. In the experiments section, we will test its performance. However, the MI-CNN is intuitively constructed based on MIL. According to aforementioned analysis, some of the instances might not be responsible for the observed classification of the bags. Such redundant or irrelevant instances may bring much disturbance to the intuitive classification. So, MI-CNN can also be used as a deep feature extractor and a more effective MIL classifier is proposed based on the deep features in Section IV. For each instance c_{ij}, we can extract the output of the second fully connected layer (FC2) as its deep feature, represented as x_{ij}. Then a video bag $\mathbf{C}_i$ can be represented as $\mathbf{X}_i = \{x_{i1}, \cdots, x_{ij}, \cdots, x_{in_i}\} \subseteq \chi$.

IV. Representative Prototype Selection Based on Self-Expressive Dictionary Learning

After obtaining the deep features of each instance, this section is to design an effective MIL classifier. How to effectively prune the useless instances and keep the discriminative and representative ones still remains a challenging problem for MIL-based objectionable video recognition. Although there are excellent strategies to select prototypes [7,39,40], MILES [39] has the difficulty of holding the increasing mapping dimension when confronting large number of instances and the strategies in [7,40] are lack of representativeness so as to missing information of the instance space. So, more progressive selection framework should be designed.

It can be observed that data has the self-expressiveness property, stating that each data point in a union of subspaces can be efficiently reconstructed by a combination of other points in the dataset [41]. Representative instances selection in this paper is just to learn those instance points which can reconstruct all instances coming from the union of subspaces. If we choose instances as dictionary atoms, we can select the representative ones with proper constraint.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.2992276, IEEE Transactions on Circuits and Systems for Video Technology

A. Self-expressive Dictionary Learning With Low Rank and Sparse Coding (SEDL-LRSC)

Let $\mathbf{Q} = [q_1, q_2, \dots, q_M] = [x_{11}, \dots, x_{1n_1}, \dots, x_{i1}, \dots, x_{in_i}, \dots, x_{N1}, \dots, x_{Nn_N}] \in \mathbb{R}^{d\times M}, M = n_1 + n_2 + \cdots + n_N$, represent all the instances in training video bags. Then the matrix of data points $\mathbf{Q}$ can be considered as a self-expressive dictionary in which each point can be written as a linear combination of other points. For all instances, we can get the expression as:

$$\mathbf{Q} = \mathbf{QZ}\ , diag(\mathbf{Z}) = 0 \tag{5}$$

where $\mathbf{Z} = [z_{i,j}]_{i,j=1}^{M}$ is the reconstruction coefficient vector and the constraint z_{ii} =0 eliminates the trivial solution of writing a point as a linear combination of itself. However, the representation of q_i in the dictionary $\mathbf{Q}$ is not unique in general. This comes from the fact that the number of data points in a subspace is often greater than its dimension. So, appropriate constraint term should be added.

As we all know, some instances in video bags are redundant and useless, so self-expressive dictionary learning should have the ability of select the most informal instances. Sparse constraint can well ensure such purpose. Furthermore, instances in video bags are diverse. They may be horrible, bloody, violent, sad, nauseating, exciting and so on. We can suppose that these instances are from different subspaces. It has been proven that, under mild conditions, low-rank representation can correctly preserve the membership of samples that belong to the same subspace [42,43]. In addition, the data are often noisy and even grossly corrupted in real applications, e.g. damaged frame or noisy voice, so we add a noise term $\mathbf{E}$ to the reconstruction term $\mathbf{Q} = \mathbf{QZ}$ in Eq.(5) as $\mathbf{Q} = \mathbf{QZ} + \mathbf{E}$. Suppose that a fraction of random entries in the instance vectors are grossly corrupted, then $\mathbf{E}$ should be sparse. So, we extend the self-expressive dictionary learning with following constraints:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \beta\ \|\mathbf{Z}\|_{2,1} + \lambda\|\mathbf{E}\|_1 \tag{6}$$
$$\text{s.t.}\ \ \mathbf{Q} = \mathbf{QZ} + \mathbf{E}, \mathbf{Z} \geq 0$$

where $\|\cdot\|_*$ is the nuclear norm of a matrix [44], i.e., the sum of the singular values of the matrix, $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$-norm of a matrix and $\|\cdot\|_1$ is the ℓ_1-norm of a matrix. A nonnegative constraint is also added in Eq. (6) because it is more consistent with the biological modeling of visual data and often leads to better performance for data representation [45,46,47].

B. Optimization for SEDL-LRSC

To solve the SEDL_LRSC problem in Eq. (6), we extend the Augmented Lagrangian Multiplier (ALM) [48] that has been widely used for the standard low-rank problem. By introducing a new parameter $\mathbf{H}$, Eq. (6) can be converted into the following equivalent problem:

$$\min_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \beta\ \|\mathbf{H}\|_{2,1} + \lambda\|\mathbf{E}\|_1 \tag{7}$$
$$\text{s.t.}\ \ \mathbf{Q} = \mathbf{QZ} + \mathbf{E}, \mathbf{Z} = \mathbf{H}, \mathbf{H} \geq 0$$

The Eq. (7) can be solved by the ALM method that minimizes the following augmented Lagrange function [49,50]:

$$\begin{aligned} &\mathrm{L}(\mathbf{Z},\mathbf{H},\mathbf{E},\mathbf{Y}_1,\mathbf{Y}_2,\mu) = \|\mathbf{Z}\|_* + \beta\|\mathbf{H}\|_{2,1} + \\ &\lambda\|\mathbf{E}\|_1 + \langle \mathbf{Y}_1, \mathbf{Q} - \mathbf{QZ} - \mathbf{E}\rangle + \langle \mathbf{Y}_2, \mathbf{Z} - \mathbf{H}\rangle + \\ &\tfrac{\mu}{2}(\|\mathbf{Q} - \mathbf{QZ} - \mathbf{E}\|_F^2 + \|\mathbf{Z} - \mathbf{H}\|_F^2) \\ &= \|\mathbf{Z}\|_* + \beta\|\mathbf{H}\|_{2,1} + \lambda\|\mathbf{E}\|_1 + \tfrac{\mu}{2}\Big(\Big\|\mathbf{Q} - \mathbf{QZ} - \mathbf{E} + \\ &\tfrac{\mathbf{Y}_1}{\mu}\Big\|_F^2 + \Big\|\mathbf{Z} - \mathbf{H} + \tfrac{\mathbf{Y}_2}{\mu}\Big\|_F^2\Big) - \tfrac{1}{2\mu}(\|\mathbf{Y}_1\|_F^2 + \|\mathbf{Y}_2\|_F^2) \end{aligned} \tag{8}$$

where $<\cdot>$ indicates inner product, $\mathbf{Y}_1$ and $\mathbf{Y}_2$ are Lagrange multipliers, and $\mu > 0$ is a penalty parameter. The function (8) is minimized by updating each of the variables $\mathbf{Z}$, $\mathbf{H}$ and $\mathbf{E}$ one at a time and then the Lagrange multipliers $\mathbf{Y}_1$ and $\mathbf{Y}_2$. More details represent at Appendix.

C. Representative Prototype Selection

To get more discriminative features used in bag-level feature mapping, we should select most representative instances to form the mapping prototype set. As Eq. (6) learned, the nonzero rows of $\mathbf{Z}$ indicate the representatives. Furthermore, the ℓ_2-norm of each row in $\mathbf{Z}$ provides its relative importance ranking. More precisely, representative whose corresponding row in the optimal coefficient matrix $\mathbf{Z}$ has higher ℓ_2-norm has a higher important ranking to take part in the reconstruction of many points in the instance space. Denote the jth column of $\mathbf{Z}$ as z_j, the ith row as z^i. Thus, we rank $\mathscr{s}$ representatives $p_{t_1}, \cdots, p_{t_{\mathscr{s}}}$ as $t_1 \geq t_2 \geq \cdots \geq t_{\mathscr{s}}$, i.e., p_{t_1} has the highest rank and $p_{t_{\mathscr{s}}}$ has the lowest rank, whenever for the corresponding rows of $\mathbf{Z}$ we have

$$\|z^{t_1}\|_2 \geq \|z^{t_2}\|_2 \geq \cdots \geq \|z^{t_{\mathscr{s}}}\|_2 \tag{9}$$

where $\|z\|_2$ is the ℓ_2-norm of vector z and is regarded as a measurement of the capacity of representatives.

We order instances in $\mathbf{Q}$ according to their capacity in descending order and select top $\mathscr{b}$ instances as the mapping prototype set (MPS).

D. Bag Feature Mapping and Classification

After getting the MPS, we can get the bag-level feature through the similarity-based feature mapping, as in MILES [40]. Let $\{p^1, p^2, \cdots, p^{\mathscr{b}}\}$ be the MPS. For a bag $\mathbf{X}_i$ and instance x, the similarity between them is given by

$$\mathrm{sim}(\mathbf{X}_i, x) = \max_{x_{ij}\in\mathbf{X}_i} \exp(-\theta\|x_{ij} - x\|^2)\ . \tag{10}$$

where θ is a similarity scale parameter. Then, the bag-level feature vector of $\mathbf{X}_i$ can be constructed based on the similarity between $\mathbf{X}_i$ and each instance in MPS as following:

$$f_{\mathbf{X}_i} = [\mathrm{sim}(\mathbf{X}_i, p^1), \mathrm{sim}(\mathbf{X}_i, p^2), \cdots, \mathrm{sim}(\mathbf{X}_i, p^{\mathscr{b}})]^T. \tag{11}$$

Similarly, we can get the bag-level feature of a given test video bag $(\mathbf{X}', y')$ as:

$$f_{\mathbf{X}'} = [\mathrm{sim}(\mathbf{X}', p^1), \mathrm{sim}(\mathbf{X}', p^2), \cdots, \mathrm{sim}(\mathbf{X}', p^{\mathscr{b}})]^T. \tag{12}$$

Finally, a SVM classifier is introduced to recognize whether a video bag is objectionable or not.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.2992276, IEEE Transactions on Circuits and Systems for Video Technology

V. Experiments

A. *Experiment Setup*

(i) Data Set

Due to the lack of publicly available video sets for objectionable video recognition, we collect two sets from the Internet. One set includes 800 horror and non-horror video clips (referred to as Horror Set), and the other one includes 800 violent and non-violent video clips (referred to as Violent Set). In addition, a public competition data set, violent scene detection set (VSD) 2014 [51], is also introduced to evaluate the performance of our method.

- Horror Set

We download a large number of movies from the Internet which consist of 100 horror movies and 100 non-horror movies, these movies are from China, US, Japan, South Korea and Thailand etc. The genres of the non-horror movies include comedy, action, drama and cartoon. We truncate each movie into several movie clips and each clip is treated as a bag. Then we invited 10 students in our laboratory to label each video clip as one of three categories: non-horror, a little horror and horror. We define "horror" scenes as those including serial killings, ghosts, monsters, vampires, animal killing, and irreligion so as to arouse the primary emotions of fear, horror, and terror and aren't suited a child under 12-year-old to watch. If the "horror" elements of a scene evoke week emotions of fear, horror and terror and one can let child between 12 and 18 to watch, then it can be labeled "a little horror". Others can be labeled "non-horror". Then the final label of the video clip is decided by voting. Label with vote of more than 90% will be retained, otherwise it will be discarded. Finally, 400 video clips labeled "horror" and 400 video clips labeled "non-horror" with different genres are selected from the candidate video set to compose the Horror Set.

- Violent Set

Similar to Horror Set, we also collect 100 violent movies and 100 none-violent movies from Internet. After truncating and manual labeling, 400 violent video clips and 400 non-violent video clips are finally selected to compose the Violent Set. For the annotators, we define "violent" scenes as those including fights, gun shots, explosions, and self-mutilation so as to stimulate mental impulses through showing the use of force to injure others or oneself and also aren't suited child under 12-year-old to watch. Labels with more than 90% votes are decided as the final labels.

- VSD2014

VSD2014 is a benchmark violence detection dataset and gets the validation during the 2014MediaEval Violence Scenes Detection (VSD) task [52,53]. The annotations have been created by several human assessors in a hierarchical, bottom-up manner. The training set in the dataset has 24 Hollywood movies and the testing dataset is composed by a set of 7 Hollywood movies and 86 YouTube video clips.

To make the dataset suitable for testing objectionable video recognition whose target is to identify whether a video is objectionable with video's label instead of frames', we truncate the movies into video segments and each segment is treated as a video bag. Its instances are constructed by key frames and spectrograms extracted from shots in the video bag via shot detection[35]. If there are violent shots in the video bag, it is annotated as positive bag, otherwise the bag is annotated as negative bag.

(ii) Evaluation Criteria

We used the precision (Pre), recall (Rec), and F_1-measure (F_1) to evaluate the performance of an algorithm. For each data set, given the ground truth of an objectionable video set (HS) as well as recognition results (ES) of an algorithm, the precision (Pre), recall (Rec), and F_1-measure (F_1) defined in Eq. (13) are used to evaluate the performances.

$$\text{Pre} = \frac{|HS \cap ES|}{|ES|}, \text{Rec} = \frac{|HS \cap ES|}{|HS|}, \text{F}_1 = \frac{2 \times P \times R}{|P + R|}. \quad (13)$$

B. *Experiment Results*

In the experiments, for Horror and Violent data sets, we conduct 3-fold cross validations 10 times and report the average performance. The parameters are selected by cross validation on the training set with regard to the F_1-measure in each procedure. For VSD2014, we report the experiment results using the given training and testing set and select the parameters in the same way.

(i) Comparison among Different Features

In this experiment, two types of features, deep multi-instance feature and hand-crafted feature (HCR), are compared. The deep one is the proposed multimodal deep multi-instance feature (MDMF). Hand-crafted features are ones applied to recognize horror or violent video scene in [20], [29], [30] and [52]. The features in [20] are mostly extracted based on color emotion and color harmony theories for "horror" detection which we refer as "EVA". While in [29], twenty visual, audio and motional features are tested individually for horror video recognition and five features are selected which are referred as "CAM". Features extracted in [30] and [52] are designed for violent video recognition. In [30], special attributes are learned from violent video description and corresponding visual, audio and motional features are extracted. In [52], except normal visual, audio and motional features, improved trajectory features are extracted. We refer features in [30] and [52] as "Attr" and "VAT". The details of HCR are shown in Table 1.

The proposed MILRPS is used to compare the performance of different features as shown in Table 2. Comparing among the four hand-crafted features, EVA and CAM achieve better performance on horror set for their emotional features. While on three violent sets which have more action, gunshot, explosion and bloodiness scene, Attr and VAT achieve better performance because they focus on trajectory-based features, motion intensity, flame and bleeding features and so on. Even on the violent video sets, the performance comparison between

Attr and VAT indicates that the hand-crafted feature is lack of flexibility and portability on varying datasets. However, the proposed deep feature MDMF can learn the high level semantic and adapt to different data sets automatically as the best results shown in Table 2.

TABLE 1 DESCRIPTIONS OF HCR

HCR	*Feature description*
EVA [20]	Emotional intensity; Color harmony; Variance of color; Lighting; Texture; Rhythm; MFCC; Spectral power; Spectral centroid; Time domain zero crossing rate
CAM [29]	Color structure; Audio Fundamental Frequency; Audio Signature; Audio Spectrum Spread; Motion Level
Attr [30]	Motion intensity; Flame; Bleeding; Audio energy; Energy entropy;
VAT [52]	Improved dense trajectories; histograms of oriented gradients; histograms of optical flow; motion boundary histograms; trajectory shape

TABLE 2 EXPERIMENTAL RESULTS ON DIFFERENT FEATURES (%)

Evaluation Criteria	*MDMF*	*EVA*	*CAM*	*Attr*	*VAT*
		Horror			
Pre	**89.33**	86.96	86.01	85.77	85.13
Rec	**90.12**	86.33	85.75	85.01	86.05
F_1	**89.72**	86.64	86.37	85.38	85.59
		Violent			
Pre	**92.01**	86.55	87.15	88.23	87.34
Rec	**92.39**	87.12	87.92	89.62	89.35
F_1	**92.2**	86.83	87.53	88.92	88.33
		VSD2014 (Hollywood)			
Pre	**67.93**	61.61	62.73	63.58	64.25
Rec	**74.07**	70.15	70.71	71.67	71.81
F_1	**70.87**	65.60	66.48	67.38	67.81
		VSD2014 (YouTube)			
Pre	**52.54**	43.48	44.33	45.18	46.03
Rec	**79.68**	76.13	76.98	77.67	78.52
F_1	**63.32**	55.35	56.26	57.13	58.03

(ii) Comparison among Different Recognition Methods

We compare the proposed instance selection method MILRPS with single instance methods and most prevailing MIL methods using the same deep features learned by the proposed deep MIL framework, as:

MILES [39]: this is a MIL method which implicitly provides the earliest instance selection solution through bag mapping feature selection based on ℓ_1-norm SVM classifier.

MILIS [40]: this method improves MILES which proposes to use a normalized probability density function (PDF) to model all negative instances in negative bags, and then selects the most positive (negative) instance from each positive and negative bag, respectively.

mi-Graph [54]: this method uses a graph to model the context between instances in a bag.

MI-kernel [55]: this method regards each bag as a set of feature vectors and then applies a set-based kernel directly for bag classification.

MI-SVM [56]: 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.

TABLE 3 EXPERIMENT RESULTS ON DIFFERENT SET (%)

Algorithm	*Pre*	*Rec*	*F₁*
	Horror		
MILRPS	**89.33**	**90.12**	**89.72**
MI-CNN	86.88	85.43	86.15
VCNN	83.23	86.93	85.04
MC-MI-J-SC	87.61	87.95	87.78
MILES	87.91	86.31	87.11
MILIS	88.57	86.47	87.5
MIKI	86.54	85.43	85.98
ISK	86.74	86.11	86.42
MI-kernel	83.84	84.15	83.99
mi-Graph	85.75	85.44	85.6
MI-SVM	83.32	81.58	82.44
SVM	78.11	78.91	78.51
KNN	91.2	60.1	72.45
	Violent		
MILRPS	**92.01**	**92.39**	**92.2**
MI-CNN	88.62	89.03	88.52
VCNN	85.92	89.89	87.86
MC-MI-J-SC	90.27	91.09	90.67
MILES	88.92	89.10	89.0
MILIS	89.82	89.58	89.7
MIKI	87.14	88.37	87.75
ISK	89.35	89.98	89.66
MI-kernel	87.28	88.91	88.09
mi-Graph	88.7	89.23	88.97
MI-SVM	85.59	87.24	86.41
SVM	83.01	80.16	81.56
KNN	82.58	75.98	79.14
	VSD2014(Hollywood)		
MILRPS	**67.93**	**74.07**	**70.87**
MI-CNN	63.11	66.83	64.92
VCNN	61.01	67.74	64.20
MC-MI-J-SC	66.69	73.17	69.77
MILES	63.77	68.89	66.23
MILIS	63.0	70.96	66.74
MIKI	63.94	68.65	66.21
ISK	63.74	69.18	66.35
MI-kernel	61.4	66.3	63.76
mi-Graph	62.81	68.31	65.45
MI-SVM	58.81	65.17	61.83
SVM	58.14	55.92	57.01
KNN	54.65	50.87	52.69
	VSD2014(YouTube86)		
MILRPS	52.54	79.68	63.32
MI-CNN	49.17	66.92	56.68
VCNN	47.85	67.69	56.07
MC-MI-J-SC	51.41	78.61	62.16
MILES	48.34	69.61	57.06
MILIS	48.73	73.45	58.59
MIKI	**53.89**	**81.22**	**64.79**
ISK	45.87	69.53	55.27
MI-kernel	43.68	66.94	52.86
mi-Graph	45.39	69.19	54.82
MI-SVM	41.73	64.79	50.76
SVM	40.84	65.77	50.39
KNN	39.98	63.07	48.94

Multi-perspective cost-sensitive MI-J-SC (MC-MI-J-SC) [18]: this is an objectionable video recognition method based on Multi-instance joint sparse coding which considers multi-perspective of objectionable video simultaneously.

MIL with Key Instance Shift (MIKI) [57]: this method addresses the problem that the distribution of key instances varies between training and testing phase by proposing an embedding based method MIKI.

MIL based on Isolation Set-Kernel (ISK) [58]: this method investigates a novel data-dependent kernel derived directly from data and introduces it into MIL.

Video classification with CNN (VCNN) [59]: this work provides an extensive empirical evaluation of CNNs on large-scale video classification. It studies multiple approaches for extending the connectivity of a CNN in time domain to take advantage of local spatio-temporal information. In this experiment, the “frame with slow fusion” framework is selected to be compared for its good performance.

Table 3 shows the average Precision (Pre), Recall (Rec) and F$_1$-measure (F$_1$) of different methods. From Table 3, the following points can be observed:

- MILRPS achieves better performance than most other methods. This shows that reasonable mapping prototype selection could improve the objectionable video recognition accuracy.
- The results of classification with MI-CNN show that MI-CNN achieves better performance than simple MIL methods such as MI-kernel, MI-SVM and video-based methods (SVM and KNN). While the improved MIL methods based on instance selection such as MILRPS, MILES and MILIS outperform MI-CNN. It reconfirms that instance selection is an effective way to improve the recognition performance of MIL-based methods. In addition, the performance on Pre and F$_1$ of MI-CNN is better than VCNN. It is because that in VCNN the label of frame or clip is treated the same as its video. It is not exact for objectionable video in which not all frames are objectionable. While in MI-CNN such problem can be well solved by the additional “NoisyOR” layer with which the deep network can be learned with bag’s(video’s) label instead of instances’(frames’) and the higher semantic from instance to bag can be learned.
- Comparison among several MIL methods based on selected prototype mapping, including MILRPS, MILES and MILIS, shows that our method MILRPS has best performance. In MILES, the selected prototypes are not given out explicitly; instead all the instances participate in the bag feature mapping and the selection will be implicitly conducted through feature selection based on ℓ_1 -norm SVM classifier. Although it has comparable accuracy, the high dimensionality of feature mapping limits its efficiency in applications with a large number of instances, e.g. video classification. MILIS first selects prototypes whose number is fixed as bag number and then gets bag mapping feature based on the prototypes. Although it avoids the high dimensional feature mapping, it still has following problems. (1) Although the most positive (negative) instances in bags are very typical training samples, they cannot represent the distribution of positive (negative) instances very well. So they have less effect on the real classification hyper plane. (2) The instance prototypes are devoid of representativeness and sensitive to noise or corruption. Our method not only avoids high dimension’s feature mapping, but also avoids fixing one instance for each bag with sparse constraint. The low rank constraint can simultaneously consider the subspace structure of instances to confirm the more informative of prototypes.
- In MI-kernel, bag feature is designed as average similarity distance between all instances in any two bags. It can be considered as an approximate method with MILES or MILIS without instance selection. The overall evaluation among MI-kernel with MILIS, MILES, MILRPS shows that instances selection can improve the discrimination of video bag feature. More reasonable selection strategy can get better recognition result.
- Although MC-MI-J-SC has no instance selection, they also achieve good performance owing to its multi-perspective and instance context in the bag.
- Comparing MIL-based methods with SVM and KNN, we can find that MIL-based methods (frame-based methods) achieve better performance for taking the structure of video into account. Furthermore, if more cues existed in objectionable video recognition can be considered, much better performance can be achieved, just as MILRPS and MC-MI-J-SC shown.
- MIKI achieves best performance on YouTube86, because YouTube86 is a generalization set in VSD2014 with different type of videos between train and test set. MIKI just focuses to address the problem that the distribution of key instances varies between training and testing phase through learning the importance weights for transformed bag vectors and incorporating original instance weights into them to narrow the gap between train/test distributions. The comparable performance achieved by our method shows its robustness when confronting variable data because of its ability to get the principal semantic by prototype selection.

(iii) Experiments on the Representativeness of the Selected Instances

In order to validate the representativeness of selected prototypes in MILRPS, we set the total number of the selected prototypes in MILRPS to be $k_P\%$ of the instance numbers, i.e. $\mathscr{b} = k_P\% \times M$. We try different value of k_P from {10, 20, ... , 100}. For each value of k_P, the average F$_1$-measure of 10 times 3-fold cross validations on the three benchmark sets are given in Fig. 3. It is very interesting to notice that the MILRPS generally achieves stable and best performance when $k_P \in [30,40]$. This phenomenon shows that the selected instances can represent the most information of the whole instance space.

In Fig. 4 we show some selected prototype examples from two objectionable video clips in horror and violent set. Each video clip is shown with its key frames and the prototypes as frames inside the red rectangles. Note that the prototypes obtained by our algorithm capture the main events of the video and the unimportant instances are certainly pruned. Through computing the mapping distance to these prototypes, more discrimination video bag feature is obtained.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.2992276, IEEE Transactions on Circuits and Systems for Video Technology

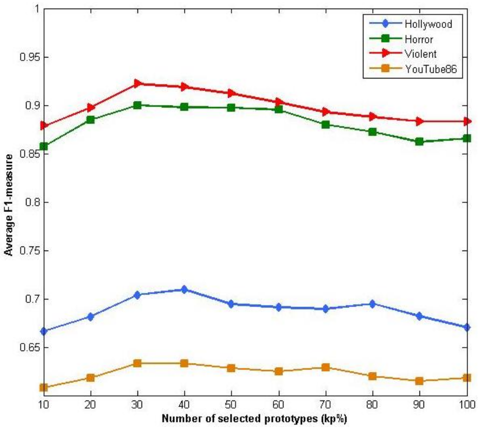

Fig.3 Average F_1-measure trained from different selected prototype number

VI. Conclusion

Two potential problems that hinder the effectiveness of MIL-based objectionable video recognition come from weak features and redundant and noise instances which contribute little to discrimination. In this paper, we proposed a MIL model based on representative prototype selection embedding deep instance representation learning. In this model, deep instance features are extracted from a novel deep network designed for MIL and a set of prototypes are selected through learning a self-expressive dictionary with sparse and low rank constraint. These prototypes not only have the representativeness of whole instances, but have the approaching distribution of instance subspaces. So, mapping to these prototypes, more discriminative bag feature can be obtained and deservedly improve the performance of objectionable video recognition. In the future, we will extend the deep multi-instance representation learning to more general classification task.

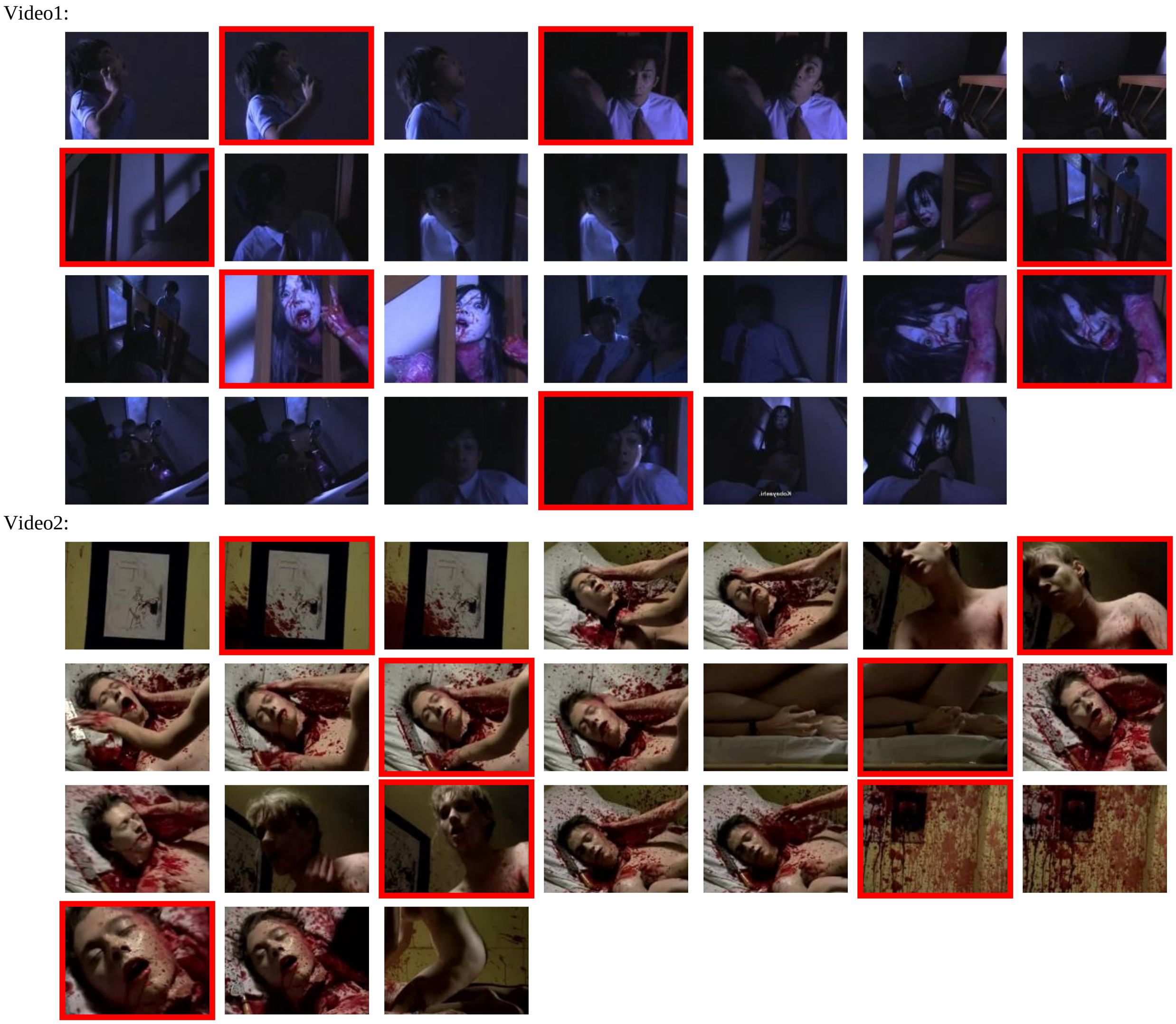

Fig 4 Examples of selected prototype

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2020.2992276, IEEE Transactions on Circuits and Systems for Video Technology

References

[1] W. Hu, O. Wu and Z. Chen et al.(2007). “Recognition of pornographic Web pages by classifying texts and images,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 29(6): 1019-1034.

[2] M. Hammami, Y. Chahir, and L. Chen(2006). “Web Guard: A Web filtering engine combining textual, structural, and visual content-based analysis,” IEEE Transactions on Knowledge Data Engineering, 18(2): 272-284.

[3] W. Hu, H. Zuo and O. Wu et al. (2011). “Recognition of adult images, videos, and Web page bags” ACM Transactions on Multimedia Computing, Communications and Applications, 7S(1): 28.

[4] P. Y. Lee, S. C. Hui, and A. C. M. Fong (2005). “An intelligent categorization engine for bilingual web content filtering,” IEEE Transactions on Multimedia, 7(6): 1183-1190.

[5] J. Tang, S. Chang and G. Qi et al.(2017). “LEGO-MM: LEarning Structured Model by Probabilistic logic Ontology Tree for MultiMedia” IEEE Transactions on Image Processing, 26(1):196-207

[6] [6] A. Araujo and B. Girod (2018). “Large-Scale Video Retrieval Using Image Queries”, IEEE Transactions on Circuits and Systems for Video Technology, 28(6): 1406-1420.

[7] J. Tang, H. Li and G. Qi.(2010) et al. “Image Annotation by Graph-Based Inference with Integrated Multiple/Single Instance Representations”. IEEE Transactions on Multimedia, 12(2): 131-141.

[8] J. Tang, J. Lin, and Z. Li, et al.(2018), “Discriminative Deep Quantization Hashing for Face Image Retrieval”, IEEE Transactions on Neural Networks and Learning Systems, 29(12): 6154 – 6162.

[9] N. Chesneau, K. Alahari and C. Schmid (2018), “Learning From Web Videos for Event Classification” , IEEE Transactions on Circuits and Systems for Video Technology, 28(10): 3019-3029.

[10] Z. Li, J. Tang and T. Mei(2019), “Deep Collaborative Embedding for Social Image Understanding”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9): 2070 – 2083.

[11] J. Tang, S. Chang and G. Qi, et al.(2017), “LEGO-MM: LEarning Structured Model by Probabilistic loGic Ontology Tree for MultiMedia”, IEEE Transactions on Image Processing, 26(1): 979-984.

[12] J. Tang, X. Shu and Z. Li, et al.(2019), “Social Anchor-Unit Graph Regularized Tensor Completion for Large-Scale Image Retagging”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(8): 2027 – 2034.

[13] J. Ye, G. Qi and N. Zhuang et al.(2018). “Learning Compact Features for Human Activity Recognition via Probabilistic First-Take-All,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(1):126-139.

[14] Y. Zhuang, J. Song and F. Wu et al. (2018) “Multimodal Deep Embedding via Hierarchical Grounded Compositional Semantics”, IEEE Transactions on Circuits and Systems for Video Technology, 28(1): 76-89.

[15] J. Tang and Z. Li, (2018), “Weakly Supervised Multimodal Hashing for Scalable Social Image Retrieval”, IEEE Transactions on Circuits and Systems for Video Technology, 28(10): 2730 – 2741.

[16] S. Qian, T. Zhang and C. Xu, et al. (2016). “Multi-Modal Event Topic Model for Social Event Analysis” IEEE Transactions on Multimedia, 18(2): 233-246.

[17] K. Li, G. Qi and J. Ye et al. (2017). “Linear Subspace Ranking Hashing for Cross-modal Retrieval.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(9):1825-1838.

[18] W. Hu, X. Ding and B. Li et al.(2016). “Multi-Perspective Cost-Sensitive Context-Aware Multi-Instance Sparse Coding and Its Application to Sensitive Video Recognition”, IEEE Transactions on Multimedia, 18(1):76-89.

[19] P. Yang, W. Song and G. S. Yang et al.(2014). “Research on Horror Video Recognition Algorithms”, Applied Mechanics and Materials, 556-562:4077-4080.

[20] J. C. Wang, B. Liand W.M. Hu et al.(2010), “Horror movie scene recognition based on emotional perception”. in Proceedings of IEEE International conference on Image Processing, 1489-1492.

[21] A. Datta, M. Shah and N. da Vitoria Lobo(2002), “Person-on-person violence detection in video data”, in Proceedings of 16th International Conference on Pattern Recognition, 1: 433-438.

[22] J. Nam, M. Alghoniemy and A.H. Tewfik(1998), “Audio-visual content-based violent scene characterization”, in Proceedings of IEEE International Conference on Image Processing, 1: 353-357.

[23] F.D.M. de Souza, G.C. Chávez and E.A. do Valle et al.(2010), “Violence Detection in Video Using Spatio-Temporal Features”, in Proceedings of 23rd SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), 224-230.

[24] E. Acar, F. Hopfgartner and S. Albayrak(2013). “Detecting violent content in Hollywood movies by mid-level audio representations”, in Proceedings of 11th International Workshop on Content-Based Multimedia Indexing (CBMI), 73-78.

[25] G. Shinichi and A. Terumasa(2015). “Violent Scenes Detection based on Automatically-generated Mid-level Violent Concepts”, in Proceedings of 19th Computer Vision Winter Workshop, 1-8.

[26] J. Tang, X. Hua and G. Qi et al.(2007). “Typicality ranking via semi-supervised multiple-instance learning.” Proceedings of the 15th ACM International Conference on Multimedia. Doi:10.1145/ 1291233.1291296.

[27] G. Qi, X. Hua and Y. Rui et al.(2007). “Concurrent multiple instance learning for image categorization.” IEEE Conference on Computer Vision and Pattern Recognition. Doi: 10.1109/CVPR.2007.383152.

[28] J.C.Wang, B.Li, and W.M.Hu et al.(2011). “Horror video scene recognition via multiple-instance learning,” in in Proceedings of International Conference on Acoustics, Speech and Signal Processing, 1325-1328.

[29] B. Wu, X. Jiang, and T. Sun et al.(2011). “A novel horror scene detection scheme on revised multiple instance learning model,” in Proceedings of International Conference on Multi Media Modeling, 6524: 359-370.

[30] S. Hao, O. Wu and W. Hu et al.(2013). “Multi-Modal Multiple-Instance Learning and Attribute Discovery with the Application to the Web Violent Video Detection”. in Proceedings of 4th International Conference on Intelligence Science and Big Data Engineering, 8261: 449-456.

[31] Y. Jiang, Z. Wu and J. Tang, et al.(2018). “Modeling Multimodal Clues in a Hybrid Deep Learning Framework for Video Classification”, IEEE Transactions on Multimedia, 20(11): 3137 – 3147.

[32] Y. Jiang, Z. Wu and J. Wang, et al.(2018). “Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(2): 352-364.

[33] W. Liu, T. Mei and Y. Zhang et al.(2015). “Multi-task deep visual-semantic embedding for video thumbnail selection” in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1:3707-3715.

[34] Guo-Jun Qi (2019). “Loss-Sensitive Generative Adversarial Networks on Lipschitz Densities.” International Journal of Computer Vision, https://xs.scihub.ltd/https://doi.org/ 10.1007/s11263 -019-01265-2.

[35] Z. Cernekova, I. Pitas, and C. Nikou(2006), “Information theory-based shot cut/fade detection and video summarization,” IEEE Transactions on Circuits and Systems for Video Technology , 16(1): 82-91.

[36] B. E. D. Kingsbury, N. Morgan, S. Greenberg (1998). “Robust speech recognition using the modulation spectrogram” . Speech Communication, 25(1-3):117-132.

[37] A. Krizhevsky, I. Sutskever and G. E. Hinton (2012). “ImageNet Classification with Deep Convolutional Neural Networks”, in Proceedings of International Conference on Neural Information Processing Systems.

[38] O. Maron and T. Lozano-Perez(1998). “A framework for multiple-instance learning”. in Proceedings of Conference on Neural Information Processing Systems, 10: 570-576.

[39] Y. Chen, J. Bi, and J. Wang(2006). “Miles: Multiple-instance learning via embedded instance selection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 28: 1931-1947.

[40] Z. Fu, A. Robles-Kelly, and J. Zhou(2011). “Milis: Multiple instance learning with instance selection.” IEEE Transactions on Pattern Analysis and Machine Intelligence, 33: 958-977.

[41] E. Elhamifar and R. Vidal(2013). “Sparse subspace clustering: Algorithm, theory, and applications”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(11):2765-2781.

[42] G. Liu, Z. Lin, and Y. Yu(2010). “Robust subspace segmentation by low-rank representation,” in Proceedings of 27th International Conference on Machine Learning, 663-670.

[43] G. Liu, Z. Lin and S. Yan et al.(2013). “Robust recovery of subspace structures by low-rank representation,” IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(1): 171-184.

[44] J.-F. Cai, E. J. Cand és, and Z. Shen(2010). “A singular value thresholding algorithm for matrix completion,” SIAM Journal on Control and Optimization, 20(4): 1956-1982.

[45] P. O. Hoyer(2003). "Modeling receptive fields with non-negative sparse coding," Neurocomputing, 52–54:547–552.

[46] D. D. Lee and H. S. Seung(1999). "Learning the parts of objects by nonnegative matrix factorization," Nature, 401(6755): 788-791.

[47] D. D. Lee and H. S. Seung(2001). "Algorithms for non-negative matrix factorization," in Advances in Neural Information Processing Systems 13. Cambridge, MA, USA: MIT Press, 556-562.

[48] Z. Lin, M. Chen, and Y. Ma(2010). "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices". UIUC Technical Report UILU-ENG-09-2214.

[49] L. Zhuang, S. Gao and J. Tang et al.(2015). "Constructing a Nonnegative Low-Rank and Sparse Graph with Data-Adaptive Features", Transactions on Image Processing, 24(11): 3717-3727.

[50] Y. Zhang, Z. Jiang and LS Davis(2013). "Learning Structured Low-rank Representations for Image Classification", in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 676-683.

[51] M. Schedl, M. Sj öberg and I. Mironicǎ et al.(2015). "VSD2014: A dataset for violent scenes detection in Hollywood movies and web videos," in Proceedings of International Workshop Content-Based Multimedia Indexing, Jun, 1–6. VSD task URL: http://www. Multimedia eval.org/mediaeval2014/violence2014.

[52] M. Sj öberg, B. Ionescu and Y. Jiang et al. (2014). "The MediaEval 2014 Affect Task: Violent Scenes Detection", Working Notes in Proceedings of MediaEval 2014: Multimedia Benchmark Workshop, Barcelona, Spain, October.

[53] Y. Jiang, Q. and C. Tan et al.(2012). "The Shanghai-HongKong Team at MediaEval2012: Violent Scene Detection Using Trajectory-based Features", in Proceedings of MediaEval 2012 Workshop.

[54] Z. Zhou, Y. Sun, and Y. Li(2009). "Multi-instance learning by treating instances as non-I.I.D. samples," in Proceedings of International conference on Machine Learning, 1249-1256.

[55] T. G ärtner, P. A. Flach and A. Kowalczyk et al.(2002). "Multi-instance kernels," in Proceedings of International conference on Machine Learning, 179-186.

[56] S. Andrews, I. Tsochantaridis, and T. Hofmann (2003). "Support vector machines for multiple instance learning," in Proceedings of Conference on Neural Information Processing Systems, 561-568.

[57] Y. Zhang and Z. Zhou (2017). "Multi-instance learning with key instance shift". in Proceedings of the 26th International Joint Conference on Artificial Intelligence, 3441-3447.

[58] B. Xu, K. Ting, and Z. Zhou (2019). "Isolation set-kernel and its application to multi-instance learning". in Proceedings of the 25th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 941-949.

[59] A. Karpathy , G. Toderici and S. Shetty et al. (2014). "Large-Scale Video Classification with Convolutional Neural Networks" in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 1725-1732.

[60] Z. Lin, R. Liu, and Z. Su(2011). "Linearized alternating direction method with adaptive penalty for low-rank representation," in Advances in Neural Information Processing Systems 24. Red Hook, NY, USA: Curran Associates, 612-620.

[61] J. Cai, E. Candes, and Z. Shen(2010). "A singular value thresholding algorithm for matrix completion." SIAM Journal of Optimization, 20(4):1956-1982.

APPENDICES

Optimization Detail of SEDL-LRSC

Let $g(\mathbf{Z},\ \mathbf{H},\ \mathbf{E},\ \mathbf{Y}_1,\ \mathbf{Y}_2,\ \mu)\ =\ \frac{\mu}{2}\ (\left\|\mathbf{Q}-\mathbf{QZ}-\mathbf{E}+\frac{\mathbf{Y_1}}{\mu}\right\|_F^2+\left\|\mathbf{Z}-\mathbf{H}+\frac{\mathbf{Y_2}}{\mu}\right\|_F^2)$, Eq.(8) is simplified as following:

$$\mathrm{L}(\mathbf{Z},\mathbf{H},\mathbf{E},\mathbf{Y}_1,\mathbf{Y}_2,\mu)=\|\mathbf{Z}\|_*+\beta\|\mathbf{H}\|_{2,1}+\boldsymbol{\lambda}\|\mathbf{E}\|_1+g(\mathbf{Z},\mathbf{H},\mathbf{E},\mathbf{Y}_1,\mathbf{Y}_2,\mu)-\frac{1}{2\mu}(\|\mathbf{Y}_1\|_F^2+\|\mathbf{Y}_2\|_F^2). \tag{14}$$

We replace the quadratic term g by its first order approximation and add a proximal term [60]; With some algebra, the updating schemes are as follows, in which the subscript k indicates the iteration number.

- Updating $\mathbf{Z}_{k+1}$:

$$\mathbf{Z}_{k+1}=\operatorname{argmin}_{\mathbf{Z}}\|\mathbf{Z}\|_*+\langle\nabla_{\mathbf{Z}}g(\mathbf{Z}_k,\mathbf{H}_k,\mathbf{E}_k,\mathbf{Y}_{1,k},\mathbf{Y}_{2,k},\mu_k),\mathbf{Z}-\mathbf{Z}_k\rangle+\frac{\eta_1\mu_k}{2}\|\mathbf{Z}-\mathbf{Z}_k\|_F^2=\underset{\mathbf{Z}}{\operatorname{argmin}}\|\mathbf{Z}\|_*+\frac{\eta_1\mu_k}{2}\left\|\mathbf{Z}-\mathbf{Z}_k+\frac{\left[-\mathbf{Q}^T\left(\mathbf{Q}-\mathbf{QZ}_k-\mathbf{E}_k+\frac{\mathbf{Y}_{1,k}}{\mu_k}\right)+\left(\mathbf{Z}_k-\mathbf{H}_k+\frac{\mathbf{Y}_{2,k}}{\mu_k}\right)\right]}{\eta_1}\right\|_F^2=\underset{\mathbf{Z}}{\operatorname{argmin}}\ \varepsilon\|\mathbf{Z}\|_*+\frac{1}{2}\|\mathbf{Z}-\mathbf{R}\|_F^2 \tag{15}$$

where $\eta_1=\|\mathbf{Q}\|_2^2$, $\varepsilon=1/\eta_1\mu_k$ and $\mathbf{R}=\mathbf{Z}_\mathrm{k}+\frac{\left[\mathbf{Q}^\mathrm{T}\left(\mathbf{Q}-\mathbf{QZ}_\mathrm{k}-\mathbf{E}_\mathrm{k}+\frac{\mathbf{Y}_{1,k}}{\mu_k}\right)-\left(\mathbf{Z}_k-\mathbf{H}_k+\frac{\mathbf{Y}_{2,k}}{\mu_k}\right)\right]}{\eta_1}$. As suggested by [61], The solution to the above problem can be solved as:

$$\mathbf{Z}_{k+1}=\mathbf{US}_\varepsilon\mathbf{V}^T=\mathbf{U}\mathrm{T}_\varepsilon[\mathbf{S}]\mathbf{V}^T \tag{16}$$
$$\text{where }(\mathbf{U},\mathbf{S},\mathbf{V}^T)=\mathcal{SVD}(\mathbf{R})$$

Here $\mathcal{SVD}(\cdot)$ is the singular value decomposition(SVD), $\mathbf{S}$ is the singular value matrix of $\mathbf{R}$. The operator $\mathrm{T}_\varepsilon[\mathbf{S}]$ in Eq.(16) is defined by element-wise thresholding of $\mathbf{S}$, i.e., $\mathrm{T}_\varepsilon[\mathbf{S}]=\operatorname{diag}([t_\varepsilon[s_1],t_\varepsilon[s_2],\dots\dots,t_\varepsilon[s_r])$ for rank of $\mathbf{S}$ being r, and each $t_\varepsilon[s]$ is determined as:

$$t_\varepsilon[s]=\begin{cases}s-\varepsilon & \text{if } s>\varepsilon\\ s+\varepsilon & \text{if } s<-\varepsilon\\ 0 & \text{otherwise}\end{cases} \tag{17}$$

- Updating $\mathbf{H}_{k+1}$:

$$\mathbf{H}_{k+1}=\operatorname{argmin}_{\mathbf{H}\geq 0}\beta\|\mathbf{H}\|_{2,1}+\frac{\mu_k}{2}\left\|\mathbf{Z}_{k+1}-\mathbf{H}+\frac{\mathbf{Y}_{2,k}}{\mu_k}\right\|_F^2 \tag{18}$$

Set $\mathbf{B}_k=\mathbf{Z}_{k+1}+\frac{\mathbf{Y}_{2,k}}{\mu_k}$, $[\mathbf{B}_k]_l$ and $\mathbf{H}_l$ denote the l^{th} row of the matrix $\mathbf{B}_k$ and $\mathbf{H}$, rewrite (18) as:

$$\mathbf{H}_{k+1}=\operatorname{argmin}_{\mathbf{H}_1,\mathbf{H}_2,\dots,\mathbf{H}_M}\sum_{l=1}^M(\beta\|\mathbf{H}_l\|_2+\frac{\mu_k}{2}\|[\mathbf{B}_k]_l-\mathbf{H}_l\|_2^2) \tag{19}$$

The optimization of (19) can be decomposed into M separate sub-problems. For each sub-problem, we have:

$$\operatorname{argmin}_{\mathbf{H}_l}\beta\|\mathbf{H}_l\|_2+\frac{\mu_k}{2}\|[\mathbf{B}_k]_l-\mathbf{H}_l\|_2^2 \tag{20}$$

It is easy to show that the optimal solution $\mathbf{H}_l^*$ of Eq.(20) must lie on the same direction of $[\mathbf{B}_k]_l$ and takes the form: $\mathbf{H}_l^*=\gamma[\mathbf{B}_k]_l$ with $\gamma\geq 0$. By forming the Lagrangian dual form, the analytical solution of (20) can be easily obtained:

$$\mathbf{H}_l^*=\begin{cases}(1-\frac{\beta}{\mu_k\|[\mathbf{B}_k]_l\|_2}) & \|[\mathbf{B}_k]_l\|_2>\frac{\beta}{\mu_k}\\ 0 & \|[\mathbf{B}_k]_l\|_2\leq\frac{\beta}{\mu_k}\end{cases} \tag{21}$$

The solution of (19) can be obtained by stacking $\mathbf{H}_l^*$ as:

$$\mathbf{H}_{k+1}=[\mathbf{H}_1^*,\mathbf{H}_2^*,\dots,\mathbf{H}_N^*]^T. \tag{22}$$

- Updating $\mathbf{E}_{k+1}$:

$$\begin{aligned}\mathbf{E}_{k+1}&=\underset{\mathbf{E}}{\operatorname{argmin}}\ \lambda\|\mathbf{E}\|_1+\frac{\mu_k}{2}\left\|\mathbf{Q}-\mathbf{QZ}_{k+1}-\mathbf{E}+\frac{\mathbf{Y}_{1,k}}{\mu_k}\right\|_F^2\\ &=\underset{\mathbf{E}}{\operatorname{argmin}}\ \delta\|\mathbf{E}\|_1+\frac{1}{2}\|\mathbf{E}-\mathbf{G}\|_F^2\end{aligned} \tag{23}$$

where $\delta={}^{\lambda}/_{\mu_k}$ and $\mathbf{G}=\mathbf{Q}-\mathbf{QZ}_{k+1}+\frac{\mathbf{Y}_{1,k}}{\mu_k}$. The solution to the above problem can be solved as [61]:

$$\mathbf{E}_{k+1}=\mathbf{S}_\delta=\mathrm{T}_\delta[\mathbf{S}],\ \text{where }(\mathbf{U},\mathbf{S},\mathbf{V}^T)=\mathcal{SVD}(\mathbf{G}) \tag{24}$$

The inexact solution method for SEDL_LRSC is summarized in Algorithm 1.

Algorithm 1 Solution for SEDL_LRSC

```
Input: Instance matrix $\mathbf{Q}$, parameter $\lambda$ and $\beta$.
Initialize: $\mathbf{Z}_0 = \mathbf{H}_0 = \mathbf{E}_0 = \mathbf{Y}_{1,0} = \mathbf{Y}_{2,0} = \mathbf{0}$, $\mu_0 = 0.1$,
        $\rho_0 = 1.1$, $\eta_1 = \|\mathbf{Q}\|_2^2$, $\varepsilon = 10^{-2}$, $k = 0$.
while not converged do
    Update the variables $\mathbf{Z}_{k+1}, \mathbf{H}_{k+1}, \mathbf{E}_{k+1}$ as Eq.(16), (22) and (24) respectively.
    Update Lagrang multipliers as follows:
        $\mathbf{Y}_{1,k+1} = \mathbf{Y}_{1,k} + \mu_k(\mathbf{Q} - \mathbf{Q}\mathbf{Z}_{k+1} - \mathbf{E}_{k+1})$.
        $\mathbf{Y}_{2,k+1} = \mathbf{Y}_{2,k} + \mu_k(\mathbf{Z}_{k+1} - \mathbf{H}_{k+1})$
    Update $\mu$ as follows:
        $\mu_{k+1} = \rho\mu_k$
        $\rho = \begin{cases} \rho_0, & \text{if } \frac{\mu_k \max(\sqrt{\eta_1}\|\mathbf{Z}_k - \mathbf{Z}_{k-1}\|_F, \|\mathbf{H}_k - \mathbf{H}_{k-1}\|_F, \|\mathbf{E}_k - \mathbf{E}_{k-1}\|_F)}{\|\mathbf{Q}\|_F} < \varepsilon \\ 1, & \text{otherwise} \end{cases}$
    Update $k$: $k \leftarrow k + 1$.
end while
Output: an optimal solution($\mathbf{Z}^*, \mathbf{H}^*, \mathbf{E}^*$).
```

Xinmiao Ding received the Ph.D. degree in mechanical, electronic, and information engineering from the China University of Mining and Technology, Beijing, China, in 2013. From March 2015 to March 2016, she is a Visiting Scholar with the Institute of Automation, Chinese Academy of Sciences, Beijing, China. Her main research interests include image and video analysis and understanding, machine learning, and internet security.

Bing Li received the Ph.D. degree from the Department of Computer Science and Engineering, Beijing Jiaotong University, Beijing, China, in 2009. From 2009 to 2011, he worked as a Postdoctoral Research Fellow with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing.

He is currently a Professor with CASIA. His current research interests include computer vision, color constancy, visual saliency detection, multi-instance learning, and data mining.

Yangxi Li received the Ph.D. degree from Peking University, Beijing, China. He is currently a Senior Engineer with the National Computer Network Emergency Response Technical Team/Coordination Center of China. His research interests lie primarily in multimedia search, information retrieval, and computer vision.

Wen Guo received the B.E. degree from Central South University, Changsha, China, in 2001, the M.S. degree from Shandong University, Jinan, China, and the Ph.D. degree from the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China.

He is currently an Associate Professor with Shandong Technology and Business University. His research interests include computer vision, multimedia, machine learning, and pattern recognition.

Yao Liu received the Ph.D. degree from the Department of Automation, University of Science and Technology of China, Hefei, China, in 2017. He is currently a research assistant in Beijing Institute of Applied Science and Technology. His research interests include data mining, machine learning and computer vision.

Weihua Xiong received the Ph.D. degree from the Department of Computer Science, Simon Fraser University, Canada, in 2007. His research interests include color science, computer vision, color image processing, and stereo vision.

Weiming Hu received the Ph.D. degree from the Department of Computer Science and Engineering, Zhejiang University, Zhejiang, China, in 1998. From 1998 to 2000, he was a Postdoctoral Research Fellow with the Institute of Computer Science and Technology, Peking University, Beijing.

He is currently a Professor with the Institute of Automation, Chinese Academy of Sciences (CASIA), Beijing. His research interests are visual motion analysis, recognition of web objectionable information, and network intrusion detection.