Computer Vision and Image Understanding
journal homepage: www.elsevier.com

Reliable Shot Identification for Complex Event Detection via Visual-Semantic Embedding

Minnan Luo[a], Xiaojun Chang[b,**], Chen Gong[c]

[a]*School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China.*
[b]*School of Computing Technologies, RMIT University, Melbourne, VIC 3000, Australia.*
[c]*Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technologyy, Nanjing 210094, China.*

ABSTRACT

Multimedia event detection is the task of detecting a specific event of interest in an user-generated video on websites. The most fundamental challenge facing this task lies in the enormously varying quality of the video as well as the high-level semantic abstraction of event inherently. In this paper, we decompose the video into several segments and intuitively model the task of complex event detection as a multiple instance learning problem by representing each video as a "bag" of segments in which each segment is referred to as an instance. Instead of treating the instances equally, we associate each instance with a reliability variable to indicate its importance and then select reliable instances for training. To measure the reliability of the varying instances precisely, we propose a visual-semantic guided loss by exploiting low-level feature from visual information together with instance-event similarity based high-level semantic feature. Motivated by curriculum learning, we introduce a negative elastic-net regularization term to start training the classifier with instances of high reliability and gradually taking the instances with relatively low reliability into consideration. An alternative optimization algorithm is developed to solve the proposed challenging non-convex non-smooth problem. Experimental results on standard datasets, *i.e.*, TRECVID MEDTest 2013 and TRECVID MEDTest 2014, demonstrate the effectiveness and superiority of the proposed method to the baseline algorithms.

1. Introduction

Recent years have witnessed the unprecedented booming of multimedia data generation and distribution on the Internet thanks to the growth of platforms such as YouTube, Facebook and Twitter. As a natural way by which human beings interact with these multimedia data, content-based multimedia analysis is of primary importance (Wu et al., 2000; Yuan et al., 2021; Zhang et al., 2020a; Yan et al., 2020; Zhang et al., 2020b). It therefore turns an interesting research efforts to content-based multimedia analysis for various applications such as multimedia information indexing and retrieval Cheng et al. (2019); Zhang et al. (2018), multimedia recommendation and multimedia event detection Chen et al. (2020); Zhan et al. (2019). As the first significant step in video analysis towards automatic categorization, recognition, search and retrieval, complex event detection that aims to automatically discover a particular event of interest in the videos, has attracting more and more research attention in the field of both computer vision and multimedia communities (Chang et al., 2017b; Song et al., 2017; Chang et al., 2017a; Yan et al., 2021; Ren et al., 2021).

Unlike elementary visual concept detection which focus mainly on simple actions, objects, and scenes, complex event detection is a much more challenging task for the richer content and higher level semantic abstraction of the unconstrained Internet videos. On one hand, a complex event in a video clip typically comprises of several lower level components such as multiple objects, various actions, different scenes, and the rich interactions between them; for example, *"carabiners"*, *"climbing gym"* and *"moving hands and feet along side of rock face"* can be found in the event *"rock climbing"*. On the other hand, the quality of user-generated videos in websites varies enormously; In practice many video clips contain some shots that

**Corresponding author
e-mail: minnluo@xjtu.edu.cn (Minnan Luo), cxj273@gmail.com (Xiaojun Chang), chen.gong@njust.edu.cn (Chen Gong)

are completely irrelevant to the event of interest or even misleading (Vahdat et al., 2013). This makes it difficult to model these unconstrained Internet videos precisely, and consequently could potentially devastate the performance of event detection.

Technically, the key of detecting the event of interest from multimedia data lies in feature extraction and classifier training. Since an untrimmed video lasts for a given period of time, it is usually decomposed into several shots to capture additional local information. In such a way, multiple instance learning approach proposed in (Chen et al., 2006) is intuitively used for complex event detection by representing each video as a "bag" of segments in which each segment is referred to as an instance. In the framework of multiple instance learning, a video bag is labeled positive with respect to an event of interest if at least one instance in that video is positive, while the video bag is labeled negative if all the instances in it are negative. Note that the labels are assigned only to video bags of instances, rather than the individual instance. In this sense, a positive video bag often contains some instances which are irrelevant to the event, while negative bags can also contain some instances that may appear in positive bags (Li and Vasconcelos, 2015). As a result, there are two main issues to be considered with respect to training the classifier for complex event detection:

- How to represent an instance precisely to identify its reliability?
- How to alleviate the negative effect of instances with low reliability?

To the first issue, early researches usually focus on low-lever visual features of appearance and motion in a video, such as Scale Invariant Feature Transform (SIFT) (Lowe, 2004), Laptev's Space-Time Interest Points (STIP) (Laptev, 2005), and Improved Dense Trajectory (IDT) (Wang et al., 2013; Wang and Schmid, 2014; Stein and McKenna, 2017). However, these handcrafted features are practically infeasible (Chakraborty et al., 2013). Leveraging on recent success in deep learning, convolution neural networks (CNN) features (Karpathy et al., 2014a) have been exploited and have yielded impressive performance. A complex event, however, often contains some prior knowledge, such as specific sequences or certain scenes and objects. Nevertheless, these visual information-based methods might fail to exploit such external information about the event of interest. As a result, great effort has been devoted to exploiting semantic information for multimedia event detection tasks. For example, J. SanMiguel and J. Martínez (SanMiguel and Martínez, 2012) propose a framework for complex event recognition guided by hierarchical event descriptions; Concept detectors (Snoek and Smeulders, 2010) that are typically learned from different multimedia archives, have come to be leveraged to enhance the performance because the descriptions of event often contain valuable concept information (Habibian and Snoek, 2014; Ma et al., 2013; Li et al., 2019; Mazloom et al., 2013). However, these approaches heavily depend on human knowledge to design the elementary concepts space (Fan et al., 2017; Cheny et al., 2019); Moreover, the limited number of concepts that are well-defined before training manually may result in concept mis-identification during the training process.

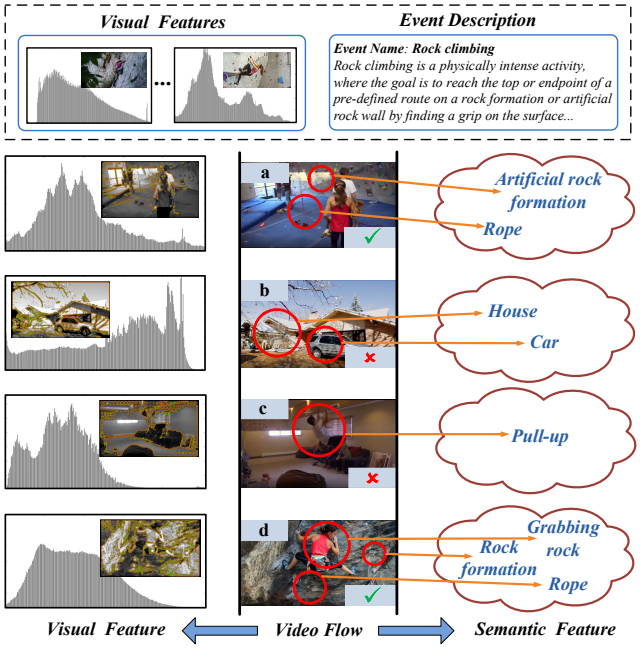

Fig. 1: An example showing the semantic and visual information of different frames.

It is noteworthy that Venugopalan *et al*. (Venugopalan et al., 2015) proposed a sequence-to-sequence model to describe visual content using natural language by mapping a video to a semantic description, and thereby achieve better performance. Although research that aims to jointly exploits visual and semantic information in video is still in its early stage, these studies have demonstrated that semantic information about a video is valuable and should not be neglected.

To the second issue on the reliability of instances, Fan *et al*. (Fan et al., 2017) select reliable instance from positive and negative video bags by inferring a binary indicator, and train classifier on the selected reliable instances only. This strategy achieves remarkable performance on multimedia event detection task. *However, this work neglects the semantic information involved in each instance and identifies the reliability according to visual information only, which might lead to incorrect results.* For example, in Figure 1, there are some instances of a rock climbing video bag. It would be quite easy to identify segments *b* and *d* as irrelevant and reliable instances respectively by looking at visual or semantic information alone. The instance *a* may be classified as irrelevant since there is no visual feature pertaining for the event *rock climbing*. However, the concepts *"artificial rock formation"* and *"rope"* found as tagged in the red circle in instance *a* are relevant to the event *rock climbing* if semantic information is taken into consideration. In this sense, instance *a* should be determined as a reliable instance and have a positive effect in the training stage. On the other hand, although instance *c* contains an action feature that resembles climbing ("pulling up"), it does not contain any semantic concept of the event *rock climbing*. Indeed, this instance may appear in many video bags of other events, such as *"physical training"* or *"indoor sport"* video bags. As a result, this instance should not be regarded as reliable in *rock climbing*

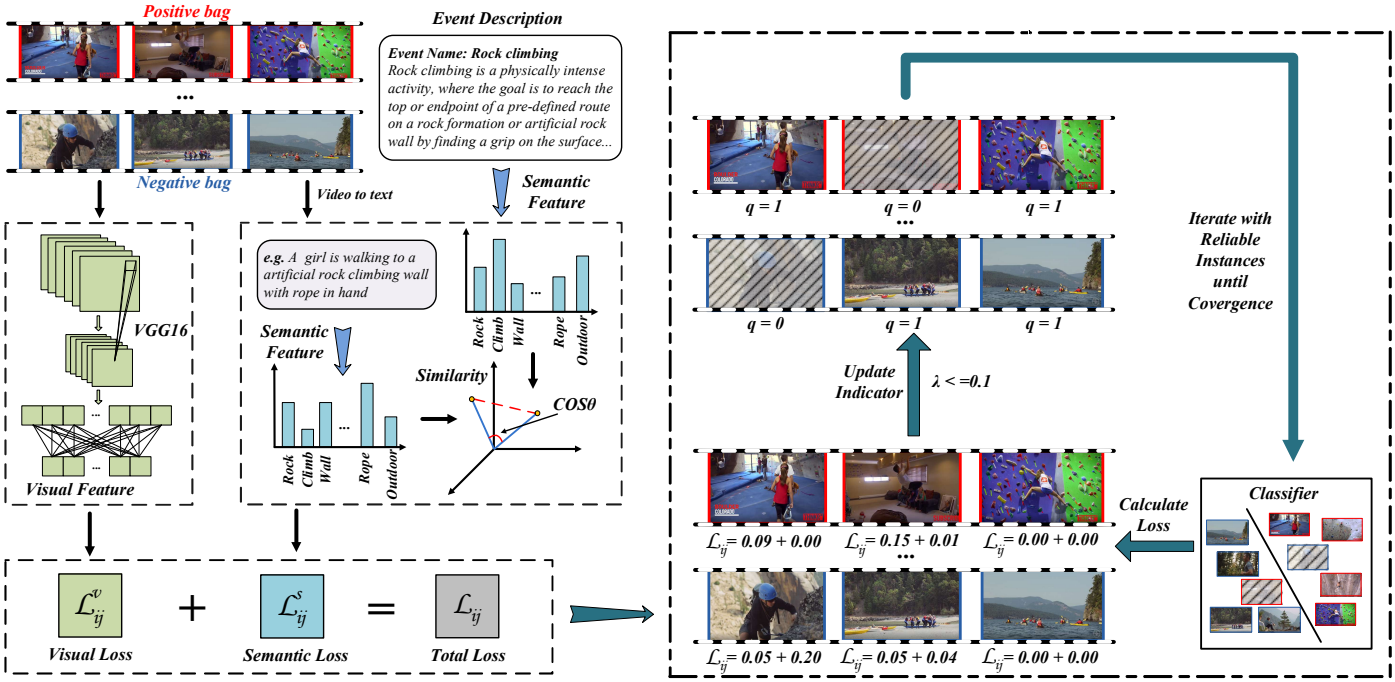

Fig. 2: The framework for our training process in multimedia event detection task.

video bag.

Figure 2 illustrates an overview of our proposed training process of multimedia event detection. In our work, we take both the high-level semantic information (similarity between the concept of event description and the semantic feature of the instance own) and low-level visual information (CNN feature of the instance) into consideration to learn a reliability variable for each instance in order to indicate its importance (see the left part of Figure 2). Since the instances with low reliability are difficult to use for training a robust classifier, we are motivated by *curriculum learning* (Bengio et al., 2009) and start training the classifier on high-reliability instances first, and then gradually take the instances with relatively low reliability into consideration (see the right part of Figure 2). We formulate our proposed approach as an optimization problem, which turns out to be highly non-convex, and hence we propose an alternating algorithm to search for the optimal value of the reliability variables and classifier parameters simultaneously. The highlights of our work are summarized as follows:

- Taking the visual low-level feature and high-level semantic information simultaneously, we propose a visual-semantic guided loss to measure the reliability of instance in the framework of multi-instance learning for event detection.
- To alleviate the negative influence of irrelevant and ambiguous segments in the training process, we begin the training with high-reliability instances and gradually added in instances with relatively low reliability over time.
- We conduct extensive experiments on two standard datasets, *i.e.*, TRECVID MEDTest 2013 and TRECVID MEDTest 2014. The promising results demonstrate the effectiveness and superiority of the proposed method to the state-of-the-art methods.

The remainder of this paper is organized as follows. In Section II, we give a brief review of some related works on multimedia event detection. Our visual-semantic guided reliable shot identification for complex event detection is proposed in Section III. An efficient algorithm is presented in Section IV for finding the solution. In Section V, extensive experiments over benchmark datasets are conducted to verify the effectiveness and superiority of the proposed algorithm. Finally, conclusions are given in Section VI.

Notations and Terms: Throughout this paper, we follow the standard notation and use normal lowercase characters for scalars (*e.g.*, $z \in \mathbb{R}$), bold lowercase characters for vectors (*e.g.*, $\mathbf{z} = [z_1, z_2, \cdots, z_d]^\top \in \mathbb{R}^d$), normal uppercase characters for matrices (*e.g.*, $Z = [\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_p] \in \mathbb{R}^{d\times p}$), and calligraphic alphabets for sets (*e.g.*, $\mathcal{Z}$). The transpose of matrix Z is denoted by $Z^\top$. The ℓ_2-norm and ℓ_1-norm of vector $\mathbf{z} \in \mathbb{R}^d$ are defined as $\|\mathbf{z}\|_2 = \sqrt{\sum_{i=1}^d z_i^2}$ and $\|\mathbf{z}\|_1 = \sum_{i=1}^d |z_i|$, respectively.

2. Related Work

In this section, we briefly review the existing related works which are relevant to multimedia complex event detection, multi-instance learning, and self-paced learning.

2.1. Multimedia Event Detection

It is crucial and difficult to represent the videos precisely for multimedia event detection since the long videos like those from TRECVID MEDTest-13 and TRECVID MEDTest-14 usually comprise of several lower level components such as multiple objects, various actions, different scenes, etc. Early

research basically extracts and aggregates various complementary low-level feature descriptors from the whole video to create a unique vector representation Zhang et al. (2017a); Ma et al. (2017); Chang and Yang (2017); Chang et al. (2016b). To name a few, Shen *et al*. (Shen et al., 2008) leverage multimodal information and apply subspace selection technique to generate video descriptor; Sun *et al*. (Sun and Nevatia, 2013) use Fisher Vector coding (Sánchez et al., 2013) as a robust feature pooling technique to combine four types of descriptors, *i.e*., motion boundary histogram (MBH) (Wang et al., 2013), histograms of gradients (HoG), optical flow (HoF) and the shape of the trajectories; Oneata *et al*. (Oneata et al., 2013) combine three feature descriptors including dense MBH (Wang et al., 2013), SIFT and mel-frequency cepstral coefficients (MFCC) audio features (Rabiner and Schafer, 2007) with Fisher vector encoding for characterizing complex event detection task, and come to a conclusion that SIFT and MFCC features provide complementary cues for complex events. Xian *et al*. (Xian et al., 2016) use a uniform experimental setup to evaluate seven different types of low-level spatio-temporal features in the context of surveillance event detection. Despite of their good performance, the low-level features fail to capture the inherent semantic information in an event.

Concept detectors that utilize several external image/video archives and learn a high-level semantic representation for the videos with complex contents (Snoek and Smeulders, 2010), are exploited to advance event detection for its consistence with human's understanding and reasoning. For example, Natarajan *et al*. (Natarajan et al., 2012) combine a large set of features from different modalities using multiple kernel learning and late score level fusion methods, where the features consists of several low-level features as well as high-level features obtained from object detector responses, automatic speech recognition, and video text recognition. Jiang *et al*. (Jiang et al., 2012) train a classifier from low-level features, encode high-level feature of concepts into graphs, and diffuse the scores on the established graph to obtain the final prediction of event. To mitigate the unavoidable noise in concept high-level features, Yan *et al*. (Yan et al., 2015) select the high-level semantic meaningful concepts based on both events-kit text descriptions and concept detectors, and learn a concept-driven event oriented dictionary representation for complex event detection; Chang *et al*. (Chang et al., 2016a) weight the semantic representations attained from different multimedia archives and propose a semantic representation analyzing framework on both source-level and the overall concept-level. Due to the growth of deep convolution neural networks (CNN) (Krizhevsky et al., 2012), CNN descriptors have been exploited for multimedia event detection and achieved impressive performance improvements (Xu et al., 2015; Zha et al., 2015). However, these traditional approaches typically extract and aggregate local descriptors from video frames or shots to create a unique vector representation for the entire video. This strategy might fail to make full use of the important structural or temporal information contained in the videos (Zhao et al., 2018), such that the key evidences are diluted for event detection, especially when the event of interest only occurs within a short period of time in an untrimmed long video.

To tackle the issues mentioned above, several research are devoted to the efforts to exploit the evidences of event interest for better performance of event detection. To name a few, Tang *et al*. (Tang et al., 2012a) divided video into several segments and discovered the discriminative and interesting segments by leaning latent variables over the frames based on the variable-duration hidden Markov model; Lai *et al*. (Lai et al., 2014b) represented each video as multiple "instances" with different temporal intervals, and inferred the instance labels and the instance-level classifier simultaneously. Fan *et al*. (Fan et al., 2017) also followed the multi-instance learning framework and estimated a linear SVM classifier together with the selection procedure of reliable training instances. Note that these approaches focus on the visual information contained in each instance (segment) and ignore the semantic information. As a result, Chang *et al*. (Chang et al., 2017b) prioritized the segments according to their semantic saliency scores which assess the relevance of each shot with the event of interest, and then developed a nearly-isotonic SVM classifiers to exploit the constructed semantic ordering information. Phan *et al*. (Phan et al., 2015) measured the importance of each segment by matching its detected concepts against the evidential description of the event interest, and jointly optimized with instance visual feature in a variant of multiple instance learning framework.

2.2. Multi-instance Learning

Multi-instance learning was first proposed in (Dietterich et al., 1997) and has been applied in several domains successfully, such as image categorization (Chen and Wang, 2004), object detection (Zhang et al., 2006), drug activation prediction (Wang et al., 2019), and retrieval (Zhang et al., 2002). In the framework of multi-instance learning, an example is regarded as an instance, while a bag labeled as positive or negative is composed of several instances. Specifically, a positive bag is defined as containing at least one positive instance, while a negative bag contains no positive instances. The classifier is finally designed to classify bags, rather than individual instances. Note that the label of a bag can be assigned easily once all instances have been labeled. It is noteworthy that Tibo *et al*. (Tibo et al., 2020) introduced multi-multi instance learning for particular way of interpreting predictions, where examples are organized as nested bags of instances. Various works have been published about bag representation by merging instances (Gärtner et al., 2002), the instance distributions of bags (Bunescu and Mooney, 2007), and the relation between multi-instance learning and semi-supervised learning (Zhou et al., 2009); however, these methods are based on the constraint that a positive bag can be determined by the existence of at least one positive instance. This assumption leads to a lack of analysis of other positive instances and is too strict for negative bag (Li et al., 2011). Since negative bags may contain several positive instances in many tasks, some works that focus on relaxing the above constraint have been developed. For example, Li *et al*. (Li et al., 2011) established a general constraint that a positive bag should contain at least a certain percentage positive instance. Moreover, considering that the bag can be represented by key instances, a clustering algorithm has been applied to detect these key instances

(Liu et al., 2012). Li and Vasconcelos (Li and Vasconcelos, 2015) showed that using the most positive-liked k instances can result in better performance. However, the instance labels are updated under weak supervision, which may lead to unreliable solutions (Zhang et al., 2015). For multimedia event detection task, a video is regarded as a bag and the segments of the video are treated as instances, after which the classifier is trained on the instances. Intuitively, the inclusion of irrelevant and ambiguous segments may have a negative influence on the training classifier. However, there has been limited research on training classifiers using only those instances with strong correlations to an event for the complex event detection tasks (Fan et al., 2017; Li et al., 2018b,a; Luo et al., 2018).

2.3. Self-pace Learning

Inspired by the learning mode of human beings, curriculum learning (Bengio et al., 2009) and self-pace learning (Kumar et al., 2010; Luo et al., 2017) were proposed to learn from easy samples to hard samples in training process to alleviate the negative effect of noisy samples. Different from curriculum learning based on certain easiness measurements, self-pace learning can automatically and dynamically choose the training order of all samples during training process (Jiang et al., 2014b). Original self-paced learning is focus on the easiness of samples, Jiang *et al*. (Jiang et al., 2014b) proposed an approach called self-paced learning with diversity which formalizes the preference for both easiness and diversity of samples via a non-convex negative $\ell_{2,1}$-norm. Self-pace learning has been widely applied to various fields, such as object tracking (Supancic and Ramanan, 2013), image classification (Tang et al., 2012b), and multimedia event detection (Jiang et al., 2014a). Self-paced learning mode can also be integrated to many existing frameworks to enhance the performance. Zhang *et al*. (Zhang et al., 2017b) combine the multi-instance learning problem with self-pace learning to improve the performance in co-saliency detection. Li *et al*. (Li et al., 2016) proposed a multi-objective method to enhance the convergence of the self-pace learning algorithms. Zhao *et al*. (Zhao et al., 2015) introduced a soft self-paced regularizer to matrix factorization to impose adaptive weights to samples. Zhou *et al*. (Zhou et al., 2018) applied self-paced learning framework into deep learning to learn the stable and discriminative features. Huang *et al*. (Huang et al., 2019) developed a similarity-aware network representation learning based on self-paced learning by accounting for both the explicit relations and implicit ones. Dizaji *et al*. (Ghasedi et al., 2019) exploited a balanced self-paced learning algorithm for deep generative adversarial clustering network.

3. The Proposed Methodology

In this section, we first introduce a visual-semantic guided loss measure at the instance level, and then propose a multi-instance learning based reliable shot identification model for multimedia event detection tasks.

In this paper, multimedia event detection task regarding event e is formulated as a binary classification problem in the framework of multi-instance learning, where the event e usually comes with a short textual description in most video event datasets such as TRECVID MED13 and TRECVID MED14. Formally, we use the skip-gram neural network model (Mikolov et al., 2013) in natural language processing to convert the textual description of event e into a vector representation $\mathbf{e} \in \mathbb{R}^d$. Suppose there are n video bags for the detection of event e, denoted by $\mathcal{B} = \{(\mathcal{B}_i, y_i) : y_i \in \{-1, 1\}; i = 1, 2, \cdots, n\}$, where $\mathcal{B}_i$ refers to the i-th untrimmed long video which are partitioned as a set of video shots (instances), *i.e*., $\mathcal{B}_i = \{I_{ij} : j = 1, 2, \cdots, m_i\}$ for $i = 1, 2, \cdots, n$. If the i-th video bag $\mathcal{B}_i$ belongs to the event e, $y_i = 1$; otherwise, $y_i = -1$. Without loss of generality, we initialize the instance label y_{ij} as the corresponding bag label, *i.e*., $y_{ij} = y_i$ for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$. We extract the CNN features from a set of uniformly sampled frames within each video shot I_{ij} to represent this video shot as $\mathbf{x}_{ij} \in \mathbb{R}^p$. To exploit the semantic information contained in the video shots, we generate a text description $\mathbf{t}_{ij} \in \mathbb{R}^d$ for each video shot I_{ij} using an end-to-end sequence-to-sequence model proposed in (Venugopalan et al., 2015).

3.1. Visual-semantic Guided Loss Measure

To make better use of the semantic and visual information of a video simultaneously, we intuitively define the visual-semantic guided loss measure on each instance I_{ij} as a convex combination of losses with regard to semantic and visual information respectively, *i.e*.,

$$\mathcal{L}_{ij} = \alpha \mathcal{L}^v_{ij} + (1 - \alpha)\mathcal{L}^s_{ij} \tag{1}$$

for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$, where $\alpha \in [0, 1]$ controls the influence of losses on visual and semantic information. Specifically, the visual information loss on the video shot I_{ij} is calculated based on hinge loss function by

$$\mathcal{L}^v_{ij}(\mathbf{w}, b; \mathbf{x}_{ij}, y_{ij}) = \max(0, 1 - y_{ij}(\mathbf{w}^T \mathbf{x}_{ij} + b)) \tag{2}$$

where $\mathbf{w}$ and b are parameters to learn.

The instances exhibiting higher correlation at the semantic level are more important to the event (Phan et al., 2015) than the ones with lower correlation. As a result, we measure the similarity between the semantic feature of each video shot I_{ij} and the event of interest e by

$$s^e_{ij} = \cos(\mathbf{t}_{ij}, \mathbf{e})$$

for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$, namely the instance-event similarity. Note that this similarity measurement is different from the one defined in (Phan et al., 2015) which is formulated based on limited number of concepts. With the instance-event similarity, the label of instance I_{ij} is predicted by function $h_r : [0, 1] \to \{-1, 1\}$ with respect to a related level threshold $r \in \mathbb{R}$ for all video shots, *i.e*.,

$$h_r(s^e_{ij}) = \begin{cases} 1, & \text{if } Rank(s^e_{ij}) \le r \\ -1, & \text{otherwise} \end{cases} \tag{3}$$

where the function $Rank(s^e_{ij})$ is utilized to quantifies the similarity s^e_{ij} into a related level. For each instance I_{ij}, the value of $Rank(s^e_{ij})$ being less than r means that the instance can be predicated as having a positive semantic similarity level with high

confidence. It is evident that this confidence increases as the value of threshold r decreases. We define the semantic information loss L^s_{ij} by penalizing the noisy instances, *i.e.*,

$$\mathcal{L}^s_{ij}\left(h_r(s^e_{ij}), y_{ij}\right) = \begin{cases} y_{ij}(1-2s^e_{ij})+1, & h_r(s^e_{ij}) \neq y_{ij} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Note that the loss of correctly labeled instance is set as 0 according to Equation (4). For the wrongly labeled instances, moreover, we calculate loss on the following cases:

- When the wrongly labeled instance I_{ij} is in a positive bag, its semantic loss $\mathcal{L}^s_{ij}$ turns to $2-2s^e_{ij}$, which means we can apply a higher penalty to those instances with lower similarity to the event of interest.
- When the wrongly labeled instance I_{ij} is in a negative bag, its semantic loss $\mathcal{L}^s_{ij} = 2s^e_{ij}$, which suggests we can apply a higher penalty to those instances with higher similarity to the event of interest.

3.2. Reliable Shots Identification in Multi-instance Learning Framework

To characterize the importance of instances with different reliability, we introduce a variable $q_{ij} \in \{0,1\}$ for each instance I_{ij}, and collect the reliability variables of video bag $\mathcal{B}_i$ into $\mathbf{q}_i = [q_{i1}, q_{i2}, \cdots, q_{im_i}]^\top \in \{0,1\}^{m_i}$ for $i = 1, 2, \cdots, n$. Intuitively, the instance I_{ij} accrues more reliability when the value of q_{ij} gets closer to 1. To identify the reliable instances for multimedia event detection task, we assign nonzero weights to reliable instances on the one hand and disperse these instances across more bags on the other hand. Following the strategy used in (Fan et al., 2017), we learn the latent reliability variables $\{\mathbf{q}_i\}_{i=1}^n$ and the classifier parameters $\mathbf{w}, b$ jointly by minimizing the weighted training loss together with the elastic-net regularization term, *i.e.*

$$\begin{aligned} \min_{\mathbf{w},b,\{\mathbf{q}_i\}_{i=1}^n} & \sum_{i=1}^{n} (\mathbf{q}_i \odot \mathcal{L}_i(\mathbf{w}, b; \alpha, r) + \Omega(\mathbf{q}_i, \lambda, \gamma)) \\ s.t. \quad & \sum_{j=1}^{m_i} q_{ij} \geq p_i m_i \ (i = 1, 2, \cdots, n) \\ & q_{ij} \in \{0,1\} \ (i = 1, 2, \cdots, n; j = 1, 2, \cdots, m_i) \end{aligned} \tag{5}$$

where "$\odot$" denotes the element-wise product; The first constraint refers that the proportion of reliable instances in each video bag is not less than $p_i \in \mathbb{R}$ $(i = 1, 2, \cdots, n)$. $\mathcal{L}_i(\mathbf{w}, b; \alpha, r) = [\mathcal{L}_{i1}, \mathcal{L}_{i2}, \cdots, \mathcal{L}_{im_i}]^\top \in \mathbb{R}^{m_i}$ represents the visual-semantic guided loss measurement on the i-th video bag, where its j-th component is calculated by

$$\mathcal{L}_{ij} = \alpha \mathcal{L}^v_{ij}(\mathbf{w}, b; \mathbf{x}_{ij}, y_{ij}) + (1-\alpha)\mathcal{L}^s_{ij}\left(h_r(s^e_{ij}), y_{ij}\right) \tag{6}$$

for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$. The regularization term $\Omega(\mathbf{q}_i, \lambda, \gamma)$ is defined as an negative elastic-net regularization term combining the l_1-norm and l_2-norm, *i.e.*,

$$\Omega(\mathbf{q}_i, \lambda, \gamma) = -\lambda\|\mathbf{q}_i\|_1 - \gamma\|\mathbf{q}_i\|_2 \tag{7}$$

for $i = 1, 2 \cdots, n$, where the parameters λ and γ are imposed on the reliability term and diversity term, respectively. Specifically, on one hand, the reliability term $\lambda\|\mathbf{q}_i\|_1$ tends to assign nonzero weights to the instances with high reliability over the instances with relatively low reliability. However, when $\lambda \neq 0$ and $\gamma = 0$, the selected instances may come from only specific bags, which may lead to overfitting. On the other hand, the diversity term $\gamma\|\mathbf{q}_i\|_2$ tends to assign nonzero weights to diverse instances residing in more bags. When $\lambda = 0$ and $\gamma \neq 0$, the algorithm selects only diverse instances so that some noisy instances may be selected, which may make the model to become biased.

4. Optimization Strategy

The objective function of optimization problem Equation (5) is non-convex and non-smooth, thus it is difficult to find the global minimum. In this section, we exploit an efficient alternative optimization algorithm to address this challenging problem.

To make the optimization problem Equation (5) tractable, we set the related level r within the range of $\{1, 2, \cdots, 10\}$ and choose the best performance. Considering the computational complexity, we initialize the instance label y_{ij} to be the same as the label of its corresponding video bag $\mathcal{B}_i$ for $i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$. The reliability variable q_{ij} is initialized as 1 for every instance I_{ij} ($i = 1, 2, \cdots, n$ and $j = 1, 2, \cdots, m_i$).

Update $(\mathbf{w}, b)$. This step aims to update the multi-task network parameters with the fixed reliability variable Q over the video set $\mathcal{B}$. Note that the instance-event similarity s^e_{ij} is fixed for each instance I_{ij}, and therefore the semantic loss $\mathcal{L}^s_{ij}$ becomes a constant after y_{ij} is fixed according to Equation (4). Moreover, the elastic-net regularization term $\Omega(\mathbf{q}_i, \lambda, \gamma)$ also becomes a constant. As a result, the optimal parameter $(\mathbf{w}, b)$ is obtained by solving the following optimization problem

$$\min_{\mathbf{w},b} \sum_{i=1}^{n}\sum_{j=1}^{m_i} q_{ij} \max(0, 1 - y_{ij}(\mathbf{w}^T\mathbf{x}_{ij} + b)) \tag{8}$$

It is evident that this problem can be easily solved as a classic weighted SVM problem through using some pre-computed kernel technique.

Update $\{\mathbf{q}_i\}_{i=1}^n$. With fixed variables $\mathbf{w}$ and b, the optimal indicator matrices $\{\mathbf{q}_i\}_{i=1}^n$ corresponding to n video bags $\mathcal{B}_i$ $(i = 1, 2, \cdots, n)$ can be learned individually. Specifically, for the i-th video bag $\mathcal{B}_i$, the visual-semantic guided loss of its j-th instance turns to a constant $\mathcal{L}_{ij}$. As a result, its reliable vector $\mathbf{q}_i$ over m_i instances in the i-th video bag $\mathcal{B}_i$ is learned by solving the following optimization problem

$$\begin{aligned} \min_{\mathbf{q}_i \in \{0,1\}^{m_i}} & \sum_{j=1}^{m_i} q_{ij}\mathcal{L}_{ij} - \lambda\|\mathbf{q}_i\|_1 - \gamma\|\mathbf{q}_i\|_2 \\ s.t. \quad & \sum_{j=1}^{m_i} q_{ij} \geq p_i m_i \end{aligned} \tag{9}$$

Algorithm 1 Optimization Procedure

Input: Instance feature$\{\mathbf{x}_{ij}\}$; Instance-similarity s^e_{ij}; Video-level label$\{Y_i\}_{i=1}^n$; Reliability ratio $\{p_i\}_{i=1}^n$; Parameters α, λ, γ, r.

Output: $\mathbf{w}$, b and $\mathbf{q}_i$ $(i = 1, 2, \cdots, n)$.

Initialize: $q_{ij} \leftarrow 1$, $y_{ij} \leftarrow y_i$ $(i = 1, 2 \cdots, n; j = 1, 2, \cdots, m_i)$.

for $r = 1$ to 10 **do**
  while not converge **do**
    Optimize $(\mathbf{w}, b)$ by SVM with fixed $\mathbf{q}_i$ $(i = 1, 2, \cdots, n)$.
    for $i = 1$ to N **do**
      Calculate loss $\mathcal{L}_i(\mathbf{w}, b, \alpha, r) \in \mathbb{R}_i^m$ according to Equation (6) and sort its components in ascending order $\rightarrow \mathcal{L}_i^{'}$.
      Set $g(t)$ as the index mapping such that $\mathcal{L}_{it}^{'} = \mathcal{L}_{ig(t)}$ $(t = 1, 2, \cdots, m_i)$.
      for $t = 1$ to m_i **do**
        if $t \leq p_i m_i$ or $\mathcal{L}_{it}^{'} < \lambda + \frac{\gamma}{\sqrt{t} + \sqrt{t-1}}$ **then**
          $q_{ig(t)} \leftarrow 1$
        else
          $q_{ig(t)} \leftarrow 0$
        end if
      end for
    end for
  end while
end for

for $i = 1, 2, \cdots, n$. We follow the algorithm proposed in (Jiang et al., 2014b) to achieve the global optimum of this non-convex problem. Specifically, we ascend the instance losses in the i-th video bag with fixed variable $\mathbf{w}, b$ and let $g(t)$ be the corresponding index mapping such that

$$\mathcal{L}_{it}^{'} := \mathcal{L}_{ig(t)} \tag{10}$$

for $t = 1, 2, \cdots, m_i$. In this sense, the $g(t)$-th instance will be selected as a reliable one for training if the following inequality

$$t \leq p_i m_i \; or \; \mathcal{L}_{it}^{'} < \lambda + \frac{\gamma}{\sqrt{t} + \sqrt{t-1}} \tag{11}$$

holds, *i.e.*, the reliability variable $q_{ig(t)}$ will be reset as 1 and vice versa. It is noteworthy that, on one hand, since the rank t has a value within the range 1 to m_i for the i-th video bag, the constraint $t \leq p_i m_i$ guarantees that at least $p_i m_i$ instances are selected as reliable ones. On the other hand, the threshold $\lambda + \frac{\gamma}{\sqrt{t} + \sqrt{t-1}}$ decreases when the rank t increases for each video bag, which can make selected instances comes from more different bags to keep diversity (Jiang et al., 2014b).

We summarize the overall algorithm for the optimization problem Equation (5) in Algorithm 1. Note that we start training the classifier using only those reliable instances with high reliability. As the iterations proceed and the training error becomes smaller, more instances will satisfy the constraint Equation (11) to minimize the loss function of optimization problem Equation (5). As a result, more and more instances will be selected into the reliable instance set to train the classifier. The computational complexity of the first step which updates the variables $\mathbf{w}, b$ is $O(pf^2)$, where $f = \sum_i^n \sum_j^{m_i} q_{ij}$ refers to the number of reliable instances used for training. During the second step which selects the reliable instances, the computational complexity of calculating the loss is $O(nm_i p)$ and the sorting process costs $O(nm_i log m_i)$; Moreover, due to the inequality $p >> log m_i$ usually holds in practice, the computational complexity of this step turns to $O(nm_i p)$ for $i = 1, 2, \cdots, n$. In summary, the computational complexity is $O\left(pf^2 + O(nm_i p)\right)$ for each iteration of Algorithm 1.

Fig. 3: Exemplars from the TRECVID MEDTest 2014 and MEDTest 2013 datasets.

5. Experiments

In this section, we conduct thorough experimental evaluations of the proposed framework. Firstly, we compare the proposed algorithm against state-of-the-art alternative baselines. We then compare it against state-of-the-art models using a single feature. After that, we compare it against state-of-the-art systems; finally, we conduct an ablation study to demonstrate the benefit of each component in the proposed algorithm.

5.1. Experimental setup

Datasets: Following existing works on Multimedia Event Detection, we evaluate the proposed algorithm on two real-world event detection datasets. These datasets have been compiled by the National Institute of Standard and Technology (NIST) for the TRECVID Multimedia Event Detection competition. To the best of our knowledge, these datasets are the largest public datasets for complex event detection.

- MEDTest14 (NIST, 2014): The TRECVID MEDTest 2014 dataset has 100 positive training examples for each event, along with about 5,000 negative samples. There are approximately 23,000 testing videos. This dataset contains events E021 to E040. Some example events are *grooming an animal*, *changing a vehicle tire*, *etc*. Please refer to (NIST, 2014, 2013) for a complete list of event names and descriptions.

- MEDTest13 (NIST, 2013): Similar to MEDTest14, there are 100 positive training examples for each event, together with about 5,000 negative samples in the MEDTest13 dataset. There are also about 23,000 testing videos. It contains events E006 to E015 and E012 to E030. A complete

list of event names and descriptions is provided in (NIST, 2013).

Table 1: Evaluating the performance of the proposed algorithm against alternative baselines. mAP is used as an evaluation metric. The performance is reported in percentages. Larger value indicates better performance.

	MED14		MED13	
	100Ex	10Ex	100Ex	10Ex
SVM	22.8	18.3	26.9	20.7
RR	23.6	18.8	27.5	21.2
Sparse MIL	19.8	14.3	23.3	16.4
SIL-SVM	22.2	18.5	24.0	19.8
SMIL-TopK	36.9	26.2	40.5	26.9
MIL-SRI	38.6	28.4	43.1	28.7
Ours	**43.2**	**31.8**	**48.7**	**33.9**

Setting: For all experiments, we strictly follow the *100Ex evaluation procedure* outlined in (NIST, 2013, 2014). Following the rules specified in the event kits, we **separately detect each event**, resulting in 20 individual tasks for each dataset. In other words, event detection is a binary classification task. We use the official split released by the NIST. For each event in the dataset, we have 100 positive training samples, and about 5,000 negative samples. Once the model has been trained, we evaluate it on the testing videos. In this paper, we consider both the *100Ex* and *10Ex* settings provided by the NIST.

Feature Extraction: We first segment each video into multiple shots using the color histogram difference as the indication of the shot boundary. In line with existing works on event detection, we simply choose the center frame from each shot, resize it to 224×224, and extract features from the $fc6$ layer of VGG16 (Simonyan and Zisserman, 2015). We run a state-of-the-art video-to-text model on each segment, then generate a description for each segment (Venugopalan et al., 2015).

Evaluation Metric: Following the NIST standard, we evaluate the event detection performance using mean Average Precision (mAP). Average Precision, which has been widely used in the area of information retrieval, is a single-valued metric approximating the area under the precision-recall curve; mAP is the mean of AP over all event classes.

5.2. Comparison against alternative baselines

In this section, we compare the performance of the proposed algorithm against state-of-the-art alternative baselines. More specifically, we conduct comparisons against the following:

- **Support Vector Machine (SVM) and Ridge Regression (RR):** SVM and RR are the commonly used classifiers in the TRECVID Multimedia Event Detection (MED) competition among the top ranked teams and existing technical reports.
- **Sparse Multi-Instance Learning (Sparse MIL):** We calculate the central point for each bag via the operation of average-pooling on all instances, based on which we aim to train a bag-level classifier.

Table 2: Performance comparison against state-of-the-art alternatives that use a **single** type of feature on the TRECVID MEDTest 2014 and MEDTest 2013 datasets. mAP is used as an evaluation metric. Performance is reported in percentages. Larger value indicates better performance.

	MED14		MED13	
	100Ex	10Ex	100Ex	10Ex
LTS (Tang et al., 2012a)	27.5	16.8	34.6	18.2
SED (Lai et al., 2014a)	29.6	18.4	36.2	20.1
DP (Li et al., 2013)	28.8	17.6	35.3	19.5
STN (Karpathy et al., 2014b)	30.4	19.8	37.1	20.4
C3D (Tran et al., 2015)	31.4	20.5	36.9	22.2
MIFS (Lan et al., 2015)	29.0	14.9	36.3	19.3
CNN-Exp (Zha et al., 2015)	29.7	–	–	–
CNN + VLAD (Xu et al., 2015)	35.7	23.2	40.3	25.6
NI-SVM (Chang et al., 2017b)	34.4	26.1	39.2	26.8
MIL-SRI (Fan et al., 2017)	38.6	28.4	43.1	28.7
Ours	**43.2**	**31.8**	**48.7**	**33.9**

- **SIL-SVM:** We first assign instances' labels as the corresponding bags' labels, then train an instance-level classifier using this information.
- **SMIL-TopK:** We first select the most confident k instances in each bag, and then train an instance-level classifier based on the selected instances. The optimal parameter k is tuned via cross-validation.
- **Multi-Instance Learning by Selecting Reliable Instances (MIL-SRI):** MIL-SRI aims to adaptively select reliable instances and does not require the inference of instance labels. We tune the parameters in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ and select the best one using cross-validation.

Experimental results are reported in Table 1; from these results, we can make the following observations. Firstly, the traditional SVM and RR can obtain promising results in both settings on the datasets of interest. Secondly, Sparse MIL and SIL-SVM achieve slightly worse performance on these datasets, and this is mainly because they did not differentiate the instances. Thirdly, we observe that SMIL-TopK and MIL-SRI significantly improve the performance of event detection on the used datasets, which indicates the benefits of exploiting reliable shots. Lastly, the algorithm proposed in this paper outperforms the other baselines by a large margin. This demonstrates the benefits of jointly exploring the semantic and visual feature for reliability learning.

5.3. Comparison under a single feature

To further validate the effectiveness of the proposed model, we compare the proposed algorithm with state-of-the-art alternatives that use a **single** type of feature. The experimental results are presented in Table 2. Note that, whenever possible, we directly quote the result from the original reference; in cases where this result was not directly available, we requested the code from the authors and ran the experiments ourselves.

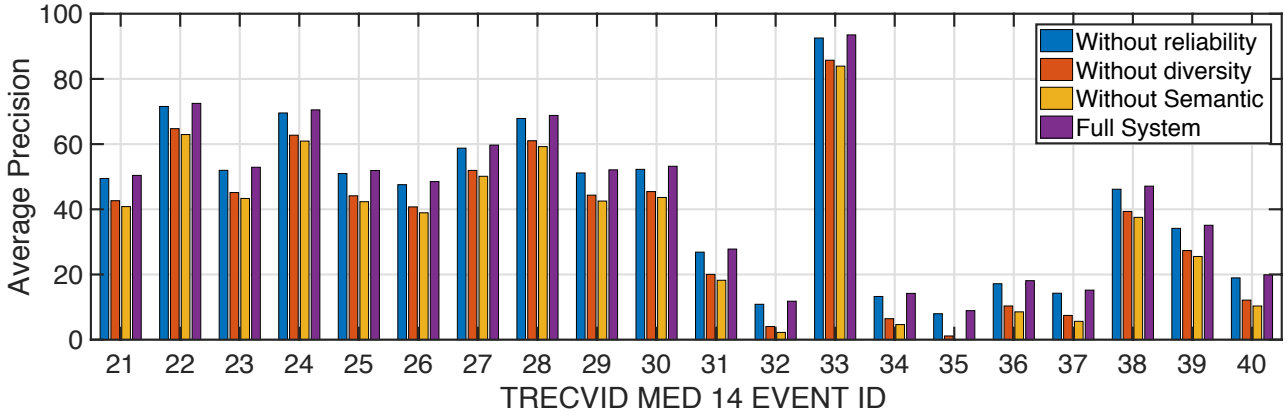

Fig. 4: We compare the results of different versions of the proposed method, including (a) without reliability; (b) without diversity; (c) without semantic; (d) full system.

Table 3: Comparison of the proposed model against state-of-the-art systems on the TRECVID MEDTest 2014 and MEDTest 2013 datasets. Performance is reported in percentages. Larger value indicates better performance.

	MED14		MED13	
	100Ex	10Ex	100Ex	10Ex
C3D (Tran et al., 2015) + IDT	33.6	22.1	39.5	26.7
CNN-Exp (Zha et al., 2015)	38.7	–	–	–
CNN + VLAD (Xu et al., 2015)	36.8	24.5	44.6	29.8
NISVM + IDT (Chang et al., 2017b)	38.1	27.2	46.3	31.5
MIL-SRI + IDT (Fan et al., 2017)	41.5	29.6	49.7	34.6
Ours + IDT	**44.9**	**33.5**	**50.3**	**35.2**

From the experimental results in Table 2, we can clearly see that the proposed algorithm outperforms the other models with a single feature by a large margin. For example, on the TRECVID MEDTest 2014 dataset, the proposed method achieves 43.2% mAP, which outperforms the second best model, MIL-SRI by 4.6% in the 100Ex setting, while also outperforming MIL-SRI by 3.4% in the 10Ex setting. This improvement is significant in the TRECVID Multimedia Event Detection competition, since event detection is a very challenging task.

5.4. Comparison with state-of-the-art systems

In the TRECVID Multimedia Event Detection competition, the top teams explore different ways to combine **multiple** different types of features. Accordingly, in this section, we also compare our method with state-of-the-art systems in the literature. In the last few years, Improved Dense Trajectories (IDT) (Wang and Schmid, 2014) have significantly outperformed the other features for the multimedia event detection competition; hence, to facilitate fair comparison, we also combine the prediction result of our method with that of IDT. The experimental results are presented in Table 3. From these results, we can see that with the proposed model, a simple combination with IDT can significantly outperform state-of-the-art systems. This further demonstrates the effectiveness of the proposed model.

5.5. Ablation study

In this section, additional experiments are conducted to confirm the effectiveness of different terms in the full system. In more detail, we compare the full system against (a) full system without reliability; (b) full system without diversity; and (c) full system without semantic. The detailed performance on an individual event is plotted in Figure 4. From the results, we can see that the full system consistently outperforms the other three alternatives; this confirms the effectiveness of all three functions. We can also observe that dropping the semantic part results in the most significant decline in performance. This phenomenon demonstrates the importance of incorporating semantic information in order to learn the reliability for each instance.

6. Conclusion

In this paper, we propose a novel approach to event detection in the framework of multi-instance learning, which learn the reliability of each instance by jointly exploiting both visual and semantic information simultaneously. To improve the robustness of the classifier, we begin the training process using high-reliability instances and gradually added in instances with relatively low reliability over time. This strategy alleviates the negative influence of irrelevant and ambiguous segments in the training process. The proposed algorithm was evaluated on two large-scale challenging datasets, and achieved very promising results. A possible direction for future work on event detection may lie in exploiting contrastive learning of visual and semantic information to extract better representations for multi-instance learning.

Acknowledgement

This work was supported by National Nature Science Foundation of China (No. 61872287 and No. 61973162), Innovative Research Group of the National Natural Science Foundation of China (No. 61721002), Innovation Research Team of Ministry of Education (IRT_17R86), Project of China Knowledge Center for Engineering Science and Technology,

NSF of Jiangsu Province (No. BZ2021013), the Fundamental Research Funds for the Central Universities (Nos: 30920032202, 30921013114), and Australian Research Council (ARC) Discovery Early Career Researcher Award (DECRA) under DE190100626.

References

Bengio, Y., Louradour, J., Collobert, R., Weston, J., 2009. Curriculum learning, in: Proceedings of the 26th International Conference on Machine Learning, pp. 41–48.

Bunescu, R.C., Mooney, R.J., 2007. Multiple instance learning for sparse positive bags, in: Proceedings of the 24th International Conference on Machine Learning, pp. 105–112.

Chakraborty, B., Gonzalez, J., Roca, F.X., 2013. Large scale continuous visual event recognition using max-margin hough transformation framework. Computer Vision and Image Understanding 117, 1356–1368.

Chang, X., Ma, Z., Lin, M., Yang, Y., Hauptmann, A.G., 2017a. Feature interaction augmented sparse learning for fast kinect motion detection. IEEE Trans. Image Process. 26, 3911–3920.

Chang, X., Ma, Z., Yang, Y., Zeng, Z., Hauptmann, A.G., 2016a. Bi-level semantic representation analysis for multimedia event detection. IEEE Transactions on Cybernetics 47, 1180–1197.

Chang, X., Yang, Y., 2017. Semisupervised feature analysis by mining correlations among multiple tasks. IEEE Trans. Neural Networks Learn. Syst. 28, 2294–2305.

Chang, X., Yu, Y., Yang, Y., Xing, E.P., 2016b. They are not equally reliable: Semantic event search using differentiated concept classifiers, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016, pp. 1884–1893.

Chang, X., Yu, Y., Yang, Y., Xing, E.P., 2017b. Semantic pooling for complex event analysis in untrimmed videos. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 1617–1632.

Chen, K., Yao, L., Zhang, D., Wang, X., Chang, X., Nie, F., 2020. A semisupervised recurrent convolutional attention model for human activity recognition. IEEE Trans. Neural Networks Learn. Syst. 31, 1747–1756.

Chen, Y., Bi, J., Wang, J.Z., 2006. Miles: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence 28, 1931–1947.

Chen, Y., Wang, J.Z., 2004. Image categorization by learning and reasoning with regions. Journal of Machine Learning Research 5, 913–939.

Cheng, Z., Chang, X., Zhu, L., Kanjirathinkal, R.C., Kankanhalli, M.S., 2019. MMALFM: explainable recommendation by leveraging reviews and images. ACM Trans. Inf. Syst. 37, 16:1–16:28.

Cheny, Z., Fuy, Y., Zhang, Y., Jiang, Y.G., Xue, X., Sigal, L., 2019. Multi-level semantic feature augmentation for one-shot learning. IEEE Transactions on Image Processing 28, 4594–4605.

Dietterich, T.G., Lathrop, R.H., Lozano-Pérez, T., 1997. Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence 89, 31–71.

Fan, H., Chang, X., Cheng, D., Yang, Y., Xu, D., Hauptmann, A.G., 2017. Complex event detection by identifying reliable shots from untrimmed videos, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 736–744.

Gärtner, T., Flach, P.A., Kowalczyk, A., Smola, A.J., 2002. Multi-instance kernels, in: Proceedings of the International Conference on Machine Learning, pp. 179–186.

Ghasedi, K., Wang, X., Deng, C., Huang, H., 2019. Balanced self-paced learning for generative adversarial clustering network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4391–4400.

Habibian, A., Snoek, C.G., 2014. Recommendations for recognizing video events by concept vocabularies. Computer Vision and Image Understanding 124, 110–122.

Huang, C., Shi, B., Zhang, X., Wu, X., Chawla, N.V., 2019. Similarity-aware network embedding with self-paced learning, in: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp. 2113–2116.

Jiang, L., Hauptmann, A.G., Xiang, G., 2012. Leveraging high-level and low-level features for multimedia event detection, in: Proceedings of the 20th ACM International Conference on Multimedia, pp. 449–458.

Jiang, L., Meng, D., Mitamura, T., Hauptmann, A.G., 2014a. Easy samples first: Self-paced reranking for zero-example multimedia search, in: Proceedings of the 22nd ACM International Conference on Multimedia, pp. 547–556.

Jiang, L., Meng, D., Yu, S.I., Lan, Z., Shan, S., Hauptmann, A., 2014b. Self-paced learning with diversity. Proceedings of the Advances in Neural Information Processing Systems , 2078–2086.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L., 2014a. Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Li, F., 2014b. Large-scale video classification with convolutional neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732.

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 1097–1105.

Kumar, M.P., Packer, B., Koller, D., 2010. Self-paced learning for latent variable models, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 1189–1197.

Lai, K.T., Liu, D., Chen, M.S., Chang, S.F., 2014a. Recognizing complex events in videos by learning key static-dynamic evidences, in: Proceedings of the European Conference on Computer Vision, pp. 675–688.

Lai, K.T., Yu, F.X., Chen, M.S., Chang, S.F., 2014b. Video event detection by inferring temporal instance labels, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2243–2250.

Lan, Z., Lin, M., Li, X., Hauptmann, A.G., Raj, B., 2015. Beyond gaussian pyramid: Multi-skip feature stacking for action recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 204–212.

Laptev, I., 2005. On space-time interest points. International Journal of Computer Vision 64, 107–123.

Li, H., Gong, M., Meng, D., Miao, Q., 2016. Multi-objective self-paced learning, in: Proceedings of the 30th AAAI Conference on Artificial Intelligence, pp. 1802–1808.

Li, W., Duan, L., Xu, D., Tsang, I.W.H., 2011. Text-based image retrieval using progressive multi-instance learning, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2049–2055.

Li, W., Vasconcelos, N., 2015. Multiple instance learning for soft bags via top instances, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4277–4285.

Li, W., Yu, Q., Divakaran, A., Vasconcelos, N., 2013. Dynamic pooling for complex event recognition, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 2728–2735.

Li, Z., Nie, F., Chang, X., Nie, L., Zhang, H., Yang, Y., 2018a. Rank-constrained spectral clustering with flexible embedding. IEEE Trans. Neural Networks Learn. Syst. 29, 6073–6082.

Li, Z., Nie, F., Chang, X., Yang, Y., Zhang, C., Sebe, N., 2018b. Dynamic affinity graph construction for spectral clustering using multiple features. IEEE Trans. Neural Networks Learn. Syst. 29, 6323–6332.

Li, Z., Yao, L., Chang, X., Zhan, K., Sun, J., Zhang, H., 2019. Zero-shot event detection via event-adaptive concept relevance mining. Pattern Recognition 88, 595–603.

Liu, G., Wu, J., Zhou, Z.H., 2012. Key instance detection in multi-instance learning, in: Proceedings of the Asian Conference on Machine Learning, pp. 253–268.

Lowe, D.G., 2004. Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60, 91–110.

Luo, M., Chang, X., Li, Z., Nie, L., Hauptmann, A.G., Zheng, Q., 2017. Simple to complex cross-modal learning to rank. Comput. Vis. Image Underst. 163, 67–77.

Luo, M., Nie, F., Chang, X., Yang, Y., Hauptmann, A.G., Zheng, Q., 2018. Adaptive unsupervised feature selection with structure regularization. IEEE Trans. Neural Networks Learn. Syst. 29, 944–956.

Ma, Z., Chang, X., Yang, Y., Sebe, N., Hauptmann, A.G., 2017. The many shades of negativity. IEEE Trans. Multim. 19, 1558–1568.

Ma, Z., Yang, Y., Xu, Z., Yan, S., Sebe, N., Hauptmann, A.G., 2013. Complex event detection via multi-source video attributes, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, pp. 2627–2633.

Mazloom, M., Gavves, E., van de Sande, K., Snoek, C., 2013. Searching in-

formative concept banks for video event detection, in: Proceedings of ACM conference on International Conference on Multimedia Retrieval, pp. 255–262.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed representations of words and phrases and their compositionality, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 3111–3119.

Natarajan, P., Wu, S., Vitaladevuni, S., Zhuang, X., Tsakalidis, S., Park, U., Prasad, R., Natarajan, P., 2012. Multimodal feature fusion for robust event detection in web videos, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1298–1305.

NIST, 2013. The trecvid med 2013 dataset. http://nist.gov/itl/iad/mig/med13.cfm.

NIST, 2014. The trecvid med 2014 dataset. http://nist.gov/itl/iad/mig/med14.cfm.

Oneata, D., Verbeek, J., Schmid, C., 2013. Action and event recognition with fisher vectors on a compact feature set, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1817–1824.

Phan, S., Le, D.D., Satoh, S., 2015. Multimedia event detection using event-driven multiple instance learning, in: Proceedings of the 23rd ACM International Conference on Multimedia, pp. 1255–1258.

Rabiner, L., Schafer, R., 2007. Introduction to digital speech processing. Foundations and Trends in Signal Processing 1, 1–194.

Ren, P., Xiao, Y., Chang, X., Huang, P., Li, Z., Chen, X., Wang, X., 2021. A comprehensive survey of neural architecture search: Challenges and solutions. ACM Comput. Surv. 54, 76:1–76:34.

Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J., 2013. Image classification with the fisher vector: Theory and practice. International Journal of Computer Vision 105, 222–245.

SanMiguel, J.C., Martínez, J.M., 2012. A semantic-based probabilistic approach for real-time video event recognition. Computer Vision and Image Understanding 116, 937–952.

Shen, J., Tao, D., Li, X., 2008. Modality mixture projections for semantic video event detection. IEEE Transactions on Circuits and Systems for Video Technology 18, 1587–1596.

Simonyan, K., Zisserman, A., 2015. Very deep convolutional networks for large-scale image recognition, in: Proceedings of the 3th International Conference on Learning Representations.

Snoek, C.G., Smeulders, A.W., 2010. Visual-concept search solved? Computer 43, 76–78.

Song, H., Wu, X., Yu, W., Jia, Y., 2017. Extracting key segments of videos for event detection by learning from web sources. IEEE Transactions on Multimedia 20, 1088–1100.

Stein, S., McKenna, S.J., 2017. Recognising complex activities with histograms of relative tracklets. Computer Vision and Image Understanding 154, 82–93.

Sun, C., Nevatia, R., 2013. Large-scale web video event classification by use of fisher vectors, in: Proceedings of IEEE Workshop on Applications of Computer Vision, pp. 15–22.

Supancic, J.S., Ramanan, D., 2013. Self-paced learning for long-term tracking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2379–2386.

Tang, K., Fei-Fei, L., Koller, D., 2012a. Learning latent temporal structure for complex event detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1250–1257.

Tang, Y., Yang, Y.B., Gao, Y., 2012b. Self-paced dictionary learning for image classification, in: Proceedings of the 20th ACM International Conference on Multimedia, pp. 833–836.

Tibo, A., Jaeger, M., Frasconi, P., 2020. Learning and interpreting multi-multi-instance learning networks. Journal of Machine Learning Research 21, 1–60.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M., 2015. Learning spatiotemporal features with 3d convolutional networks, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4489–4497.

Vahdat, A., Cannons, K., Mori, G., Oh, S., Kim, I., 2013. Compositional models for video event detection: A multiple kernel learning latent variable approach, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 1185–1192.

Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., Saenko, K., 2015. Sequence to sequence – video to text, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542.

Wang, H., Kläser, A., Schmid, C., Liu, C.L., 2013. Dense trajectories and motion boundary descriptors for action recognition. International Journal of Computer Vision 103, 60–79.

Wang, H., Schmid, C., 2014. Action recognition with improved trajectories, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 3551–3558.

Wang, X., Yan, Y., Tang, P., Liu, W., Guo, X., 2019. Bag similarity network for deep multi-instance learning. Information Sciences 504, 578–588.

Wu, J.K., Kankanhalli, M.S., Lim, J.H., Hong, D., 2000. Perspectives on Content-Based Multimedia Systems. Kluwer Academic, Hingham, MA.

Xian, Y., Rong, X., Yang, X., Tian, Y., 2016. Evaluation of low-level features for real-world surveillance event detection. IEEE Transactions on Circuits and Systems for Video Technology 27, 624–634.

Xu, Z., Yang, Y., Hauptmann, A.G., 2015. A discriminative cnn video representation for event detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1798–1807.

Yan, C., Chang, X., Luo, M., Zheng, Q., Zhang, X., Li, Z., Nie, F., 2021. Self-weighted robust LDA for multiclass classification with edge classes. ACM Trans. Intell. Syst. Technol. 12, 4:1–4:19.

Yan, C., Zheng, Q., Chang, X., Luo, M., Yeh, C., Hauptmann, A.G., 2020. Semantics-preserving graph propagation for zero-shot object detection. IEEE Trans. Image Process. 29, 8163–8176.

Yan, Y., Yang, Y., Meng, D., Liu, G., Tong, W., Hauptmann, A.G., Sebe, N., 2015. Event oriented dictionary learning for complex event detection. IEEE Transactions on Image Processing 24, 1867–1878.

Yuan, D., Chang, X., Huang, P., Liu, Q., He, Z., 2021. Self-supervised deep correlation tracking. IEEE Trans. Image Process. 30, 976–985.

Zha, S., Luisier, F., Andrews, W., Srivastava, N., Salakhutdinov, R., 2015. Exploiting image-trained cnn architectures for unconstrained video classification, in: Proceedings of the 26-th British Machine Vision Conference, pp. 1097–1105.

Zhan, K., Chang, X., Guan, J., Chen, L., Ma, Z., Yang, Y., 2019. Adaptive structure discovery for multimedia analysis using multiple features. IEEE Trans. Cybern. 49, 1826–1834.

Zhang, C., Platt, J.C., Viola, P.A., 2006. Multiple instance boosting for object detection, in: Proceedings of the Advances in Neural Information Processing Systems, pp. 1417–1424.

Zhang, D., Han, J., Jiang, L., Ye, S., Chang, X., 2017a. Revealing event saliency in unconstrained video collection. IEEE Trans. Image Process. 26, 1746–1758.

Zhang, D., Meng, D., Han, J., 2017b. Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Transactions on Pattern Analysis and Machine Intelligence 39, 865–878.

Zhang, D., Meng, D., Li, C., Jiang, L., Zhao, Q., Han, J., 2015. A self-paced multiple-instance learning framework for co-saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 594–602.

Zhang, L., Chang, X., Liu, J., Luo, M., Prakash, M., Hauptmann, A.G., 2020a. Few-shot activity recognition with cross-modal memory network. Pattern Recognit. 108, 107348.

Zhang, L., Liu, J., Luo, M., Chang, X., Zheng, Q., 2018. Deep semisupervised zero-shot learning with maximum mean discrepancy. Neural Comput. 30.

Zhang, L., Luo, M., Liu, J., Chang, X., Yang, Y., Hauptmann, A.G., 2020b. Deep top-k ranking for image-sentence matching. IEEE Trans. Multim. 22, 775–785.

Zhang, Q., Goldman, S.A., Yu, W., Fritts, J.E., 2002. Content-based image retrieval using multiple-instance learning, in: Proceedings of the Nineteenth International Conference on Machine Learning, pp. 682–689.

Zhao, Q., Meng, D., Jiang, L., Xie, Q., Xu, Z., Hauptmann, A.G., 2015. Self-paced learning for matrix factorization, in: Proceedings of the 29th AAAI Conference on Artificial Intelligence, pp. 3196–3202.

Zhao, Z., Xiang, R., Su, F., 2018. Complex event detection via attention-based video representation and classification. Multimedia Tools and Applications 77, 3209–3227.

Zhou, S., Wang, J., Meng, D., Xin, X., Li, Y., Gong, Y., Zheng, N., 2018. Deep self-paced learning for person re-identification. Pattern Recognition 76, 739–751.

Zhou, Z.H., Sun, Y.Y., Li, Y.F., 2009. Multi-instance learning by treating instances as non-iid samples, in: Proceedings of the 26th International Conference on Machine Learning, pp. 1249–1256.