IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 8, 2020, accepted May 18, 2020, date of publication May 25, 2020, date of current version June 5, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2997357

Abnormal Event Detection via Feature Expectation Subgraph Calibrating Classification in Video Surveillance Scenes

OU YE[1,2], JUN DENG[2], ZHENHUA YU[1], TAO LIU[1], AND LIHONG DONG[1]

[1]College of Computer Science & Technology, Xi'an University of Science and Technology, Xi'an 710054, China
[2]College of Safety Science and Engineering, Xi'an University of Science and Technology, Xi'an 710054, China

Corresponding authors: Jun Deng (dengj@xust.edu.cn) and Zhenhua Yu (zhenhua_yu@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61873277 and Grant 71571190, in part by the Natural Science Basic Research Plan in Shaanxi Province of China under Grant 2018JQ5095, Grant 2020JQ-758, Grant 2020JM-533, and Grant 2019JLM-11, and in part by the Key Research and Development Program of Shaanxi Province under Grant 2019GY-056 and Grant 2018ZDXM-GY-036.

ABSTRACT At present, the existing abnormal event detection models based on deep learning mainly focus on data represented by a vectorial form, which pay little attention to the impact of the internal structure characteristics of feature vector. In addition, a single classifier is difficult to ensure the accuracy of classification. In order to address the above issues, we propose an abnormal event detection hybrid modulation method via feature expectation subgraph calibrating classification in video surveillance scenes in this paper. Our main contribution is to calibrate the classification of a single classifier by constructing feature expectation subgraphs. First, we employ convolutional neural network and long short-term memory models to extract the spatiotemporal features of video frame, and then construct the feature expectation subgraph for each key frame of every video, which could be used to capture the internal sequential and topological relational characteristics of structured feature vector. Second, we project expectation subgraphs on the sparse vector to combine with a support vector classifier to calibrate the results of a linear support vector classifier. Finally, the experiments on a common dataset named UCSDped1 and a coal mining video dataset in comparison with some existing works demonstrate that the performance of the proposed method is better than several the state-of-the-art approaches.

INDEX TERMS Abnormal event detection, feature expectation subgraph, calibrating classification, sequential and topological relational characteristics.

I. INTRODUCTION

In recent years, abnormal event detection in intelligent video surveillance has gained more and more attention in academic and industrial communities [1], [2], which has become an important task in intelligent video surveillance since it is related to visual saliency [3], interestingness prediction [4], dominant behavior detection [5] and other topics in computer vision. Abnormal event detection for video sequences is a difficult challenge because of the volatility of the definitions between normality and abnormality [6] and dependence of the definitions on the context scenario. Nevertheless, it can generally be considered that abnormal behavior or an activity by unexpected events occurs less often than normal (familiar) events [7]. In order to detect abnormal events in surveillance videos, various kinds of modeling techniques are proposed in the literature, such as trajectory-based models [8], spatiotemporal feature-based models [9], [10] and sparse reconstruction-based models [11], where the majority to address the anomaly event detection are to learn and extract the hand-crafted or deep appearance features of video from given samples first and then classify and decide whether the events are abnormal if they deviate from the model of normal event.

At present, feature extraction is regarded as one key factor for abnormal event detection in existing models. From the

The associate editor coordinating the review of this manuscript and approving it for publication was Guangdong Tian.

feature representation point of view, abnormal event detection models are mainly classified into hand-crafted features-based models and deep features-based models.

In the hand-crafted features-based models, trajectory [12], flow [13] and vision modeling [14] can be used to describe the dynamic information and spatiotemporal information of video sequences, besides trajectory modeling, such as color, texture, optical flow, bag-of-words (BOW) [15] modeling and so on. The models based on color and texture features can describe appearance features in a video sequence, but they ignore motion representations. Optical flow modeling can describe dynamic information of video, but it is susceptible to illumination. The bag-of-words approach computes an unordered histogram of visual words occurrences that encode only the global distribution of low-level descriptors, but it ignores the local structural organization of salient points [16]. Although trajectory modeling can represent motion characteristic of foreground objects, it is not robust for complex scenes of video. In general, hand-crafted features-based models depend on some priori knowledge, and are not generalized well for complex video surveillance scenes.

With the development of machine learning studies, various approaches based on deep learning have achieved remarkable progress in abnormal event detection. For example, convolutional neural networks (CNNs) [17], recurrent neural networks [11] and other deep learning models can learn better feature representation than hand-crafted feature modeling. It is conductive to determinate the occurrence of abnormal event in video sequence.

Nevertheless, most of abnormal event detection models based on deep learning mainly focus on data represented in a vectorial form, which pay little attention to the impact of the internal structure characteristics of feature vector on classifying and determining abnormal events in video sequences. Moreover, a single classifier is difficult to ensure the accuracy of classification. Especially for the complex video surveillance scenes, the disturbances from the light source, occlusions and other factors in video will obviously affect the data represented in a vectorial form and accuracy of algorithm. Hence, deciding how to utilize the structure characteristics of feature vectors to filter unexpected eigenvalues that correspond to disturbances to improve the accuracy of abnormal event detection remains a challenging task.

In this paper, we propose an abnormal event detection hybrid modulation method via feature expectation subgraph calibrating classification (DF-ESCC) in video surveillance scenes to address the above issues. Our method consists of three parts: deep feature extraction, feature expectation subgraph construction and expectation subgraph-based calibration classification. First, we employ a convolutional neural network and long short-term memory (LSTM) model to extract the features in video surveillance scenes. Second, we construct the feature expectation subgraph for each key frame of every video, which could be used to capture the internal sequential and topological relational characteristics of structured feature vector. Finally, we project expectation subgraph on the sparse vector, which is used to combine support vector classifiers to calibrate the classification of linear support vector classifiers to determinate whether there exist abnormal events in the video surveillance scenes. The common dataset UCSDped1 [18] and Coal Mining video datasets [19] are used to verify the effectiveness of our proposed method. In summary, our contributions are summarized as follows: (1) we introduce the feature expectation subgraph to represent the internal sequential and topological relational characteristics of structured feature vector; (2) we propose a DF-ESCC method combining feature expectation subgraph with support vector classifiers to calibrate the classification of linear support vector classifiers; (3) the proposed method is validated on challenging UCSD dataset and coal mining video dataset, where the coal mining video dataset has complex context scenarios.

The rest of paper is organized as follows. In Section 2, we present a brief review of related abnormal event detection methods. We propose an abnormal event detection hybrid modulation method via feature expectation subgraph calibrating linear support vector classifier classification in Section 3. The experiments are performed to verify the performance of the proposed DF-ESCC method in Section 4. Finally, we conclude this paper.

II. RELATED WORK

In this section, we briefly review previous abnormal event detection models from the point of view of appearance feature for image in video. We first recall some hand-crafted features-based models, and then review deep features-based models for abnormal event detection. Finally, we analyze the shortcoming of the above methods.

A. HAND-CRAFTED FEATURES-BASED MODELS

In the past decade, trajectory features are widely used to abnormal event detection. For example, the study in [20] presents a complex event processing method based on trajectories. Song *et al.* [21] propose an approach that firstly obtains the trajectories of vehicles and pedestrians, and then detects the abnormal events using the trajectory features. However, the features of above methods are relatively single. Serhan *et al.* [22] propose to incorporate object trajectory analysis and pixel-based analysis for abnormal behavior inference and event detection, but this method is not suitable for images with poor quality. In [23], a multi-feature fusion method is used to obtain characteristic information of pedestrians, and then motion information is attained by trajectory analysis. The limitation of this method is that it is susceptible to feature changes. Moreover, the graph-based representation and learning of relevant features are combined and correlated with target behaviors to detect abnormalities in moving object trajectories, so as to determine whether the events of interest are normal or abnormal [24]. Fu *et al.* [25] utilize reference points as well as the piecewise linear segmentation algorithm to compress the trajectories, and then propose a time-aware and spatially correlated collaborative algorithm to increase the density of the trajectories to improve the accuracy of abnormal event detection. However, in this method there exists the issue of large cumulative errors in trajectory calculation. The work in [26] presents a survey of trajectory-based surveillance applications with a focus on abnormal event detection.

In general, the representation of trajectory features is sensitive to noise interference, and there exists the discontinuity of target trajectory. Thus, the models based on trajectory features are not completely reliable and not robust for the crowded scenes and other complex scenes.

To overcome the drawback of trajectory-based models, spatiotemporal features [27]–[29] are extracted from low-level appearance and motion cues to address the above problems. For example, the study in [30] proposes an approach that relies mainly on spatial abstractions of each object, mining frequent temporal patterns in a sequence of video frames to form a regular temporal pattern, which is used to detect abnormal events. However, this approach is difficult to describe spatial abstractions of each object accurately when the image features are transformed. In [31], the spatiotemporal information and slow feature analysis method are combined to represent the discriminative information in videos to detect abnormal crowd motion, but the high semantic inherent features of this method have limited ability to represent nonlinear features. The work in [32] proposes distribution of magnitude and orientation of local interest frame descriptor is used to learn a support vector machine based a binary classifier to detect violence events. Moreover, a feature descriptor is proposed by adopting the covariance matrix coding optical flow in multi-regions of interest to represent motion information, and then one-class support vector machine is applied to detect the abnormal events in [33]. The limitation of these approaches is that a single classifier is difficult to ensure the accuracy of classification. Wang *et al.* [34] propose to learn the histograms of optical flow orientations of the observed video frames by a hidden Markov model to detect abnormal events in a crowded scene. In order to at least alleviate the impact of label information on supervised or semi-supervised models, the study in [35] proposes an unsupervised algorithm that combines the manifold-based feature with a graph density search mechanism to detect abnormal network events. However, these algorithms need to know the data distribution in advance.

In summary, the models based on spatiotemporal features have a better recognition ability for moving objects with linear or nearly linear features, but they need to know prior knowledge, and have limited representation ability for nonlinear features.

B. DEEP APPEARANCE FEATURES-BASED MODELS

In order to overcome the drawback of hand-crafted features, many models based on deep appearance features are proposed to detect abnormal events. The above deep appearance features can be obtained by using convolutional neural networks [36], [37], recurrent neural networks [38], [39] and autoencoder networks [40], [41].

For example, the work in [42] first combines the saliency information with multi-scale histogram optical flow of video frames to represent spatiotemporal information, and then adopts a deep learning network named PCANet to extract high-level features of video to detect abnormal events. As an extension of the above model, Damla *et al.* [43] explore different convolutional neural networks to model patterns in a video sequence to detect abnormal behavior. In [44], the temporal convolutional neural network and optical flow models are combined to detect local anomalies. The study in [45] integrates the one-class support vector machine into convolutional neural network to implement a novel end-to-end model. The above approaches pay more attention to the extraction of spatial features, but the spatial-temporal relationship is not close enough.

In order to address that issue, it is necessary to introduce the recurrent neural networks to capture the temporal features. For example, a convolutional autoencoder integrates with a long short-term memory model to detect abnormal events in video surveillance in [46]. Kothapalli *et al.* [47] use mixture of Gaussians to subtract the background of each frame first, and then a convolutional neural network is used to extract spatial features that are fed into long short-term memory to learn temporal features. Finally, a linear support vector machine is used to classify to detect abnormal events. In [48], a novel recurrent neural network is constructed to learn sparse representation and dictionary to detect anomaly events by proposing an adaptive iterative hard-thresholding algorithm. The work in [49] combines the body shape, depth and optical flow features with long short-term memory network to implement the fall detection. The limitations of above approaches based on long-short term memory network are that the feature of noise interference will continue to spread in the process of recurrent neural network, which will affect the accuracy of features representation.

Moreover, autoencoder networks are used to detect the anomaly event. For example, an unsupervised deep feature learning algorithm is proposed by using a deep three-dimensional convolutional network and multi-level similarity trees after sparse coding to detect abnormal events in [6]. Wang *et al.* [50] use hybrid spatiotemporal autoencoder to solve the problem that long-short term memory encoder-decoder framework fails to account for the global context of the learned representation with a fixed dimension representation. The work in [51] uses a two-stream recurrent variational autoencoder to detect abnormal events in video streams. However, these approaches pay little attention to the impact of the internal structure characteristics of two feature vector on classifying and determining abnormal events in video sequences.

In summary, the models based on deep appearance feature have a better recognition ability for moving objects with nonlinear features, but the disturbance features from the light source, occlusions and other factors in video will spread in the

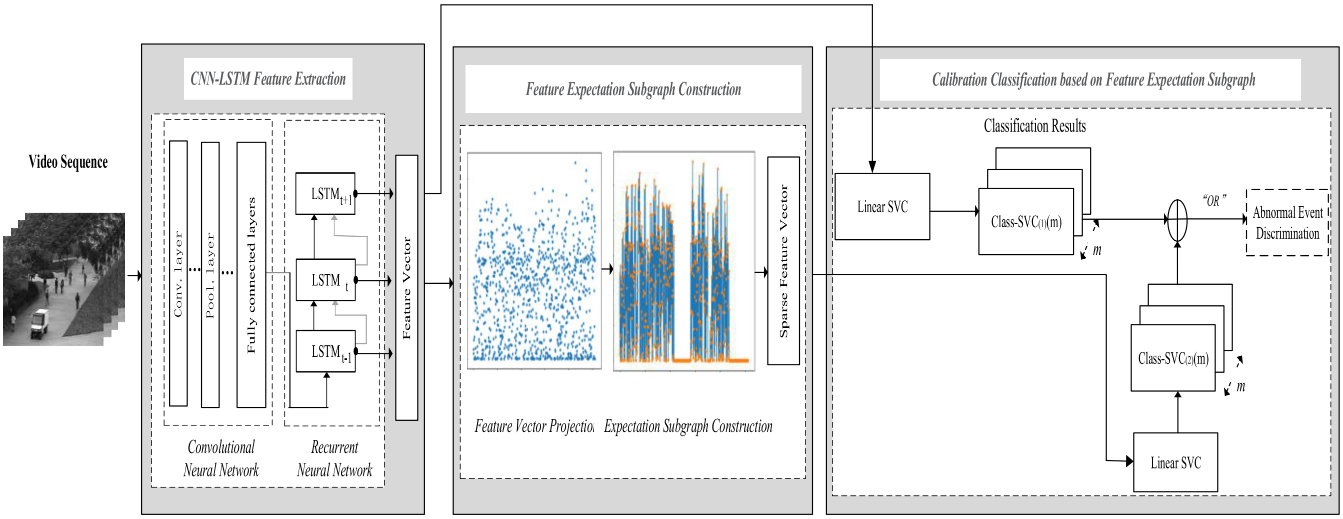


FIGURE 1. The framework of DF-ESCC method.

depth neural networks, which will seriously affect the accuracy of feature representation. In addition, a single classifier or activation function is difficult to ensure the accuracy of classification. Hence, we utilize the structure characteristics of the deep appearance features to filter unexpected feature representations, and combine a support vector classifier to calibrate the results of a single classifier.

III. THE PROPOSED METHOD

In this section, we describe how to utilize structure characteristics of the feature vector to improve the performance of abnormal event detection modeling. Recently, there have been a large number of works focusing on key-points or feature vectors to classify and detect abnormal events in video sequences. The key insight of these works is to exploit appearance feature representation and utilize probability statistical models or clustering approaches to determinate whether the events as abnormal if they deviate from the model of normal event in video surveillance scenes. However, feature representation of vectorial form is not easy to describe the topological, geometric and other complex relational characteristics of real-world data, and the disturbances from the light source, occlusions and other factors in video can affect the feature representations. Moreover, a single classifier is difficult to ensure the accuracy of classification.

In this paper, we try to construct a feature expectation subgraph to filter unexpected feature representations that from various disturbances in video and combine feature expectation subgraphs and support vector classifiers to improve the identify result of single classifier. The advantage of using feature expectation subgraphs is to obtain principal component of feature vector while retaining the sequential and topological relational characteristics inside feature vector. It is conducive to classification and recognition of abnormal event detection. Therefore, we employ the convolutional neural network and long short-term memory models to extract the features in video surveillance scenes first, and then construct expectation subgraphs by measuring the distance between eigenvalues in a feature vector. In the following, we combine expectation subgraphs with support vector classifiers to calibrate the classification of linear support vector classifiers to determinate whether there exist abnormal events in a video surveillance scene. Fig. 1 illustrates the overview of our method, which contains three parts: CNN-LSTM feature extraction, feature expectation subgraph construction and calibration classification based on feature expectation subgraph. In the following, we will describe these parts separately.

A. CNN-LSTM FEATURE EXTRACTION

Deep neural network models have more powerful learning capacity and excellent representational capacity than hand-crafted features models. Convolutional neural networks are a kind of common deep neural network, which are suitable for spatial relationships learning on raw input data. Among the various convolutional neural network models, a convolutional neural network named VGG-16 can be employed to extract spatial features as well as for high accuracy image recognition because of the depth of network [47], and therefore it can be applied to feature extraction for complex video surveillance scenes. However, the VGG-16 network is difficult to represent the temporal relationship of the input video sequences accurately. In order to overcome such a limitation, we employ a long short-term memory network to extract dynamic temporal behavior feature in video stream. In consideration of spatiotemporal features of video, we first select several video clips as the training samples to input VGG-16 network to extract the spatial features, and then the obtained feature maps are fed into LSTM to further extract the temporal features of input video clips. Suppose that the above-mentioned video clips are with a size of $w \times h \times c \times l$, where $w \times h$ denotes the size of video frame, c denotes the number of channels for each frame, and l denotes the

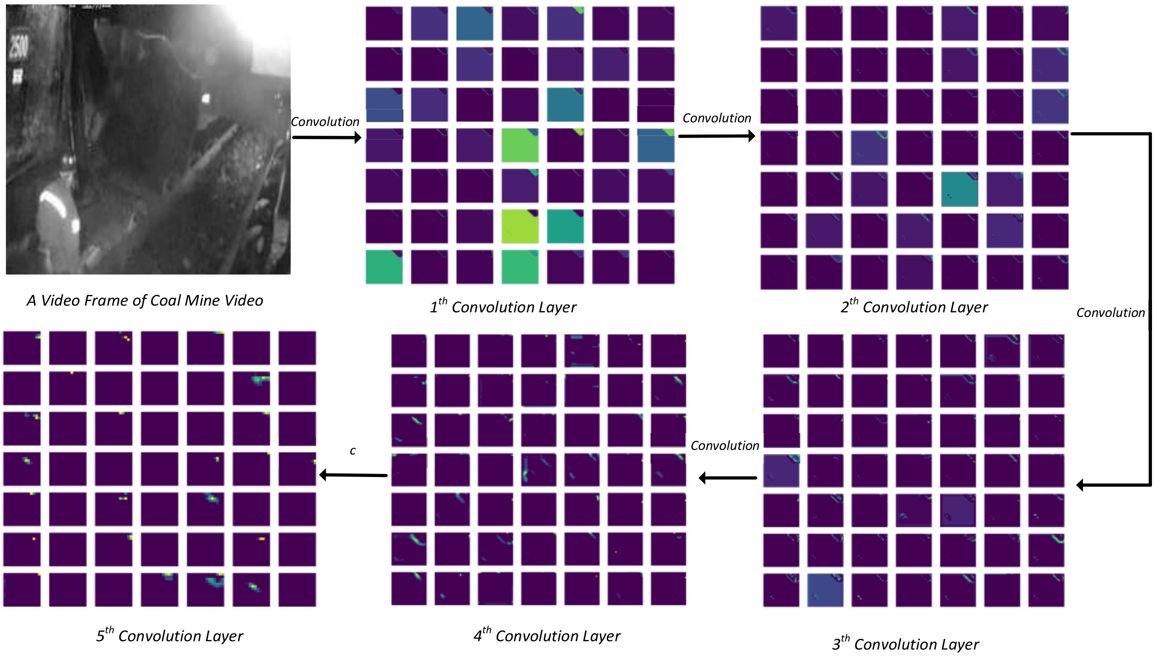


FIGURE 2. The visual feature map for different convolution layers.

frame number of the video clips. We set w and h as 224, and $c = 3$ before training the VGG-16 network. Moreover, we first fix 3×3 convolutional kernel with stride 1 in convolution layer, and then fix 2×2 pooling window with stride 2 in pooling layer to implement the convolutional operation and max-pooling process. During the process of convolutional operation, feature matrix $\mathbf{Y_{ij}}$ can be obtained by the following formulation:

$$\mathbf{Y_{ij}} = f(\mathbf{X_{ij}} \otimes \mathbf{W} + b) \tag{1}$$

where $f(\cdot)$ denotes the activation function, $\mathbf{X_{ij}}$ is the window matrix around the pixel of i^{th} row and j^{th} column in video frame, $i \in [0, h-1]$ and $j \in [0, w-1]$. Moreover, $\mathbf{W}$ denotes the weight matrix, and b is the bias. In the VGG-16 network, we select a rectified linear unit function to represent $f(\cdot)$ and set a variable z to denote the maximum value of all elements in a variable z to denote the maximum value of all elements in feature matrix $\mathbf{Y_{ij}}$, and then $f(\cdot)$ is described as follows:

$$f(z) = \max(0, z) \tag{2}$$

Through five groups of convolution and max-pooling layer, we use three fully connected layers to extract spatial feature vectors of size [4096,1]. In addition, the function of cross entropy loss is used to optimize a convolutional neural network. Taking a video frame of coal mine video as an example, we visualized the feature maps of different convolution layers, as shown in Fig. 2. It can be seen from Fig. 2 that the edge features of video frame are salient in the first convolution layer. However, with the increase of convolution layer, the feature maps are more and more abstract, and finally the high-level features of video frames are obtained.

Subsequently, the extracted feature vectors are fed into a long short-term memory network to further extract temporal feature. Here we employ a two-layer long short-term memory network, and the long short-term memory network in each layer has the same architecture, which consists of input gate, forget gate and output gate. In the process of training a long short-term memory network, we set the learning rate to 0.01, the number of input nodes to 64, and the number of nodes in hidden layer to 256. Moreover, we utilize the cross-entropy function as the loss function to train, i.e.,

$$L(y_i', y_i) = -\sum_{i=1}^{n} y_i' \times \log(y_i) \tag{3}$$

where y_i is the i^{th} eigenvalue in feature vector from output gate, y_i' denotes the label corresponding to y_i, and $i \in [1, 1024]$.

After we complete the VGG-LSTM networks training, we can obtain the feature vectors of size [1024,1] from the output layer of long short-term memory network to represent the video clips. The concrete architecture of VGG-LSTM networks is described in Fig. 3.

B. FEATURE EXPECTATION SUBGRAPH CONSTRUCTION

The disturbances from the light source, occlusions and other factors will affect feature extraction whether in normal or complex video surveillance scenes, which are also reflected in feature representations. Although the principal component analysis algorithms can reserve the main features of video frames while reducing the impact of disturbances, the structure characteristics of feature vector will change. At present, most studies mainly focus on data represented in a vectorial

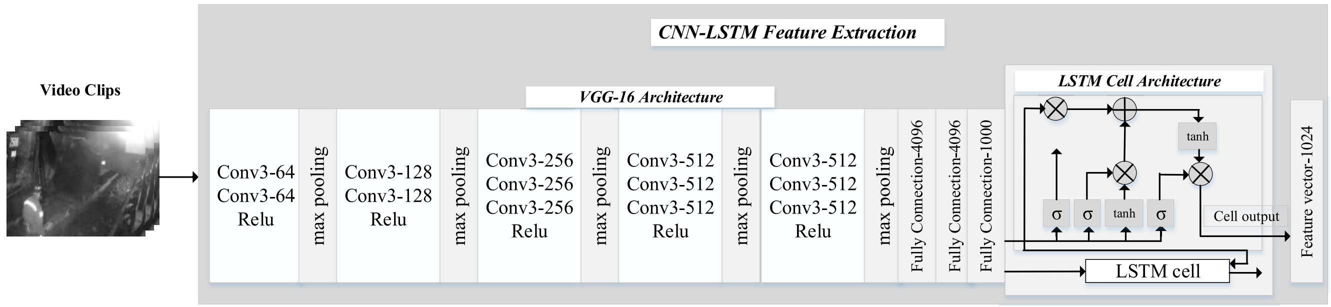


FIGURE 3. VGG-LSTM networks architecture.

form, which pay little attention to the impact of the internal structure characteristics of feature vector for abnormal event detection in video surveillance scenes. In this section, we briefly describe how to construct a feature expectation subgraph to represent sequential and topological relational characteristics between eigenvalues in a structured feature vector.

Suppose that we obtain a set of feature vectors $S = \{\boldsymbol{V}_i\}_{i=1}^{n}$ by using the VGG-LSTM networks, where the i^{th} feature vector $\boldsymbol{V}_i \in \mathbb{R}^{1024\times 1}$. Since the distribution of feature points has the sequential and topological relationships in video frame, the distance between two eigenvalues $y_n^{(i)}$ and $y_m^{(i)}$ in any one $\boldsymbol{V}_i = [y_1^{(i)}, y_2^{(i)}, \cdots, y_{1024}^{(i)}]$ is probably closer to each other if the two feature points corresponding to the $y_n^{(i)}$ and $y_m^{(i)}$ are adjacent in video frame, where m and n respectively denote the index positions of $y_m^{(i)}$ and $y_n^{(i)}$ in a feature vector, $m < n$, and $m, n \in [1, 1024]$. In order to represent the internal sequential and topological relationships of a feature vector, we first transform the feature vector to a two-dimensional matrix by using the following formulation:

$$\mathbf{A}^{(i)} = \begin{cases} y_{t,l}^{(i)}, & (t = l) \\ 0, & (t \neq l), \end{cases} \quad t, l \in [1, 1024] \tag{4}$$

where t denotes the t^{th} row, l denotes the l^{th} column in matrix $\mathbf{A}^{(i)}$, and the i^{th} matrix $\mathbf{A}^{(i)}$ corresponds to the i^{th} feature vector $\boldsymbol{V}_i = [y_1^{(i)}, y_2^{(i)}, \cdots, y_{1024}^{(i)}]$. Second, we use a mapping $\varphi : y_{t,l}^{(i)} \rightarrow P(y_{t,l}^{(i)}, l)$ to obtain an eigenvalue point in two-dimensional space if the value of an element is not 0 in $\mathbf{A}^{(i)}$. Therefore, each eigenvalue $y^{(i)}$ corresponds to an eigenvalue point $y_{t,l}^{(i)}$ in two-dimensional space. Suppose that we have two eigenvalue points $P(y_{t1,l1}^{(i)}, l_1)$ and $P(y_{t2,l2}^{(i)}, l_2)$. We can measure the distance between two eigenvalue points by using

$$dis(P(y_{t1,l1}^{(i)}, l_1), P(y_{t2,l2}^{(i)}, l_2)) = \frac{\alpha 1 \cdot \psi 1(y_{t1,l1}^{(i)}, y_{t2,l2}^{(i)}) + \alpha 2 \cdot \psi 2(l_2, l_1)}{\alpha 1 + \alpha 2} \tag{5}$$

where the parameters $t_1, t_2, l_1, l_2 \in [1, 1024]$, α_1 and α_2 are constraint factors, and $y_{t1,l1}^{(i)}, y_{t2,l2}^{(i)} \in \mathbf{A}^{(i)}$. According to [16], the position of eigenvalue points in two-dimensional space is also the main factor to measure the internal sequential and topological relationships of a feature vector, besides eigenvalue. Therefore, the first term $\psi 1(y_{t1,l1}^{(i)}, y_{t2,l2}^{(i)})$ in Eq. (5) measures the similarity of eigenvalues between two eigenvalue points, and the second term $\psi 2(l_2, l_1)$ measures the similarity of the position between two eigenvalue points. Moreover, we calculate k using Eq. (6) to roughly measure the contribution relationship between two terms for distance measurement.

$$k = \frac{\max(y_1^{(i)}, y_2^{(i)}, \cdots, y_{1024}^{(i)} \in \boldsymbol{V}_i)}{\dim(\boldsymbol{V}_i)} \tag{6}$$

where $\dim(\boldsymbol{V}_i)$ denotes the dimension of feature vector $\boldsymbol{V}_i$. On this basis, we use the Euclidean distance function to represent $\psi 1(y_{t1,l1}^{(i)}, y_{t2,l2}^{(i)})$ and $\psi_2(l_2, l_1)$; thus we can further describe Eq. (5) as below:

$$dist(P(y_{t1,l1}^{(i)}, l_1), P(y_{t2,l2}^{(i)}, l_2)) = \left\{ \alpha 1 \cdot \left(y_{t1,l1}^{(i)} - y_{t2,l2}^{(i)} \right)^2 + \alpha 2 \cdot (l_2 - l_1)^2 \right\}^{\frac{1}{2}},$$
$$s.t. \begin{cases} [0, \frac{r}{k}], & if\ K > l \\ [0, r \times k], , & if\ K > l, \end{cases} \quad \alpha_1 + \alpha_2 = 1 \tag{7}$$

where r denotes the range of neighborhood. We employ the idea of K-NearesNeighbor algorithm to calculate the distance only in r scope (we set $r = 100$ in our experiment), which can not only reduce the computational cost but also decrease the influence of the far position of eigenvalue point in feature vector on distance calculation. If, $dis(P(y_{t1,l1}^{(i)}, l_1), P(y_{t2,l2}^{(i)}, l_2)) \leq \mu_T P(y_{t1,l_1}^{(i)}, l_1)$ where μ_T is a given threshold, we consider two points $P(y_{t1,l1}^{(i)}, l_1)$ and $P(y_{t2,l2}^{(i)}, l_2)$ as similar eigenvalue points, and utilize an edge to represent incidence relation between each other. In this way, there are several eigenvalue points that will be related to each other by using edges, and several edge sets are generated to represent incidence relation of all eigenvalue points in feature vector. Through the above eigenvalue points and edge sets, we can construct a graph $G = (v, \varepsilon(v))$, where v denotes the set of eigenvalue points, and $\varepsilon(v)$ denotes the corresponding edge set. In order to utilize the structure characteristics of deep feature vectors to filter unexpected eigenvalues that correspond to disturbances to improve the accuracy of abnormal event detection, we present

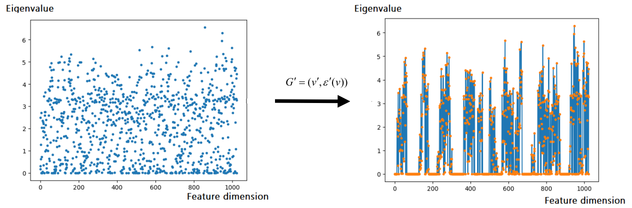


FIGURE 4. The construction of feature expectation subgraph.

to construct a feature expectation subgraph for each key frame of the video. First, we calculate the expected value of edge sets in graph G in below:

$$\mathrm{E}_{\varepsilon(v)\sim P(|\varepsilon(v)|)} = \sum_{|\varepsilon(v)|} P(\varepsilon(v))f(\varepsilon(v)) \tag{8}$$

where $f(\varepsilon(v))$ denotes discrete function on $\varepsilon(v)$. Since the probability of occurrence of any $\varepsilon(v)$ is random, we can further describe it as below:

$$\mathrm{E}_{\varepsilon(v)\sim P(|\varepsilon(v)|)} = \frac{\sum\limits_{i=1}^{N}(|\varepsilon(v)|_i)}{N} \tag{9}$$

After that, we can obtain feature expectation subgraph $G' = (v', \varepsilon'(v))$ if the condition $|\varepsilon(v)| \geq \mathrm{E}_{\varepsilon(v)\sim P(|\varepsilon(v)|)}$ is met, as shown in Fig. 4. Fig. 4(a) shows the eigenvalue points in feature vector generated from VGG-LSTM networks, and Fig. 4(b) shows a feature expectation subgraph G'. From Fig. 4(b), we can see that some eigenvalue points are filtered when some eigenvalue points do not satisfy the condition $dis(P(y^{(i)}_{t1,l1}, l_1), P(y^{(i)}_{t2,l2}, l_2)) \leq \mu_T$, and others are retained. Moreover, the graph that is composed of these eigenvalue points can reserve the main part of internal sequential and topological relational characteristics of structured feature vector. When there are less feature expectation subgraphs, we will use all feature subgraphs as feature expectation subgraphs. When a feature subgraph contains all eigenvalue points, we can regard it as the maximum feature expectation subgraph.

C. CALIBRATION CLASSIFICATION BASED ON FEATURE EXPECTATION SUBGRAPH

Once the frames of video are represented using feature expectation subgraphs, we can use them to classify and recognize anomaly. In this section, we will combine with support vector classifiers and feature expectation subgraphs to calibrate the classification of a single linear support vector classifier.

First let $\{G', y'_i\}^n_{i=1}$ be the corresponding labeled feature expectation subgraphs for n frames from N training videos $\{V_i\}^N_{i=1}$, where the label y'_i is -1 for feature expectation subgraphs of abnormal event and $+1$ for feature expectation subgraphs of normal event. Second, we utilize the support vector classifier to classify G' and detect the abnormal events. In this paper, we solve the classification problem of support vector classifier for feature expectation subgraphs, which is based on the improved support vector machine model in [16], as formulated below:

$$\min J(G'_i, G'_j, y'_i, y'_j) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y'_i y'_j K(G'_i, G'_j) - \sum_{i=1}^{n}\alpha_i,$$

$$s.t. \sum_{i=1}^{n}\alpha_i y'_i = 0 \quad \text{and } 0 \leq \alpha_i \leq C \tag{10}$$

where α_i and α_j are Lagrange multipliers, $y'_i \in \{-1, +1\}$, $K(G'_i, G'_j)$ is the graph kernel function, and C is the box constraint parameter. Since we can use an inverse mapping $\varphi' : P(y^{(i)}_{t,l}, l) \rightarrow y^{(i)}_{t,l}$ to obtain a sparse vector $\boldsymbol{V}^{(i)}_S$ that corresponds to a feature expectation subgraph G'_i, we can establish a conversion relationship: $G'_i \rightarrow \boldsymbol{V}^{(i)}_S$. On this basis, we adopt the linear kernel function $K(G'_i, G'_j) = \boldsymbol{V}^{(i)}_S \times \boldsymbol{V}^{(j)}_S$ to measure the similarity between any two G'_i and G'_j. The decision function for a test G' will be:

$$f(G', G'_i) = sign(\sum_{i=1}^{m}\alpha_i y'_i K(G'_i, G') + b) \tag{11}$$

where b is the bias, and $f(\cdot) = f(-1, +1)$ is the prediction function. Although feature expectation subgraphs can be used to obtain principal component of feature vector while reserving the main sequential and topological relational characteristics inside feature vector, a single classifier is difficult to ensure the accuracy of classification. In addition, sparse vectors obtained by feature expectation subgraphs cannot completely represent the feature of video frame. Hence, we combine with the linear support vector classifier to detect the abnormal events in video scenes as follows:

$$\hat{f}(G', G'_i, \boldsymbol{V}, \boldsymbol{V}_i) = f(\boldsymbol{V}, \boldsymbol{V}_i) \vee f(G', G'_i) \tag{12}$$

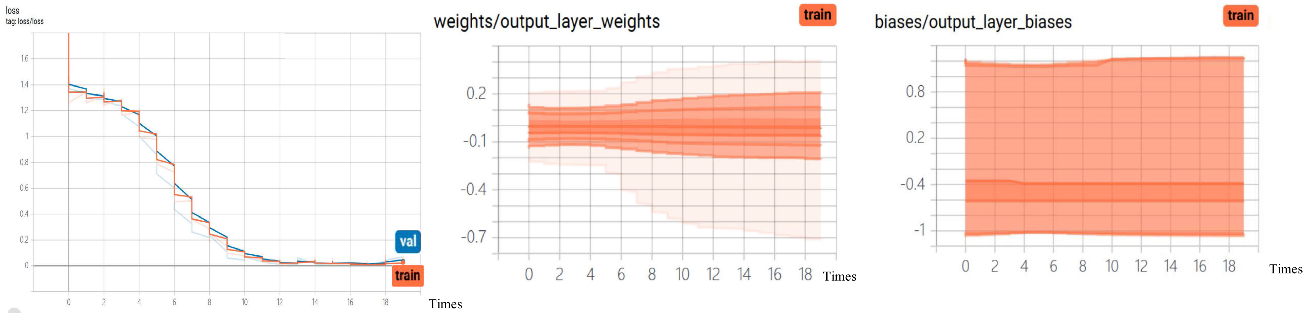
where $\boldsymbol{V}$ is the feature vector that is extracted from VGG-LSTM networks for test samples. By the logical OR operation, we can utilize the result of $f(G', G'_i)$ to calibrate the classification of $f(\boldsymbol{V}, \boldsymbol{V}_i)$.

IV. EXPERIMENTAL EVALUATION

We conduct extensive experiments on a widely used abnormal event detection dataset and a coal mining video dataset to evaluate the performance of the proposed DF-ESCC and compare them with several state-of-the-art methods such as SURF+BoW, SIFT+BoW [16], HMM with optical flow [34], CNN-2D+LSTM [43] and CNN-2D+LSTM+SVM [47]. All the experiments are conducted on a machine having a Inter Core (TM) i7-7700HQ processor with 8G memory and a Huawei server having 4-Inter Xeon processors with 8G memory, respectively. The programs are written in Python with version 3.5. In what follows, we describe the details of experiments and results.

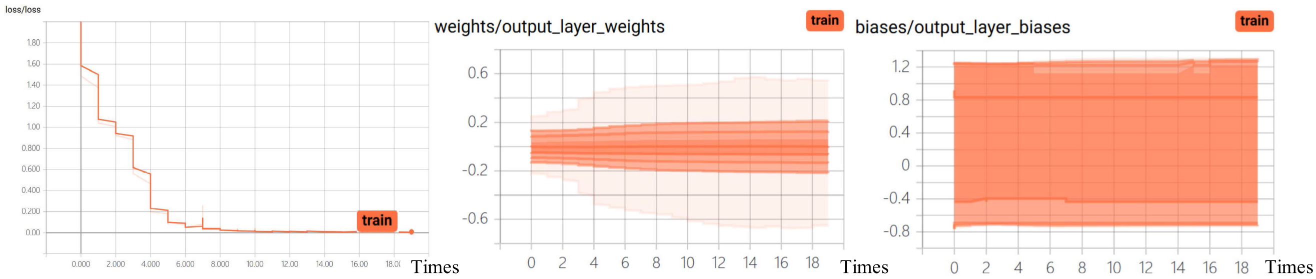
A. DATASET AND EVALUATION CRITERIA

In real life, there are video quality issues in the collected video data and more repetitive information in each video frame, which are not conducive to the detection of abnormal events in the video surveillance scenes. In order to verify



(a) Loss variation (b) weights variation in training (c) biases variation in training

FIGURE 5. The visualization of loss and weight variation during UCSDped1dataset training.



(a) Loss variation (b) weights variation in training (c) biases variation in training

FIGURE 6. The visualization of loss and weight variation during coal mining video dataset training.

the effectiveness and performance of the proposed method in common scenarios, we choose the UCSDped1 dataset [18] to evaluate the proposed method since it is the most commonly used benchmarks for abnormal event detection in videos. Moreover, we also mainly focus on the coal mine video dataset [19] as it has complex scenes, which is more challenging than UCSDped1 dataset. The abnormal event of coal accumulation is also common in production of the coal mine. By using coal mine video dataset, the validity and performance of our proposed method can be verified in complex scenarios. The UCSDped1 dataset can provide 34 training clips and 36 testing clips, and each clip has around 200 frames with a 238 × 158 pixels resolution. In our experiments, we utilize total 1393 video frames from 4 videos to detect "biker", "cart", "wheelchair" and "skater" abnormal events. In the coal mining video dataset, there are 73 videos that can be used for training and testing, and each frame is the 3-channel image of 224 × 224 pixels resolution. In our experiments, we select total 6879 frames from 6 videos in 3 scenes to detect an abnormal event of coal accumulation.

In order to evaluate the performance of the proposed approach, the following metrics [43]: accuracy, precision and recall metrics are used for evaluation of abnormal event detection, which are expressed by:

$$Precision = TP/(TP + FP) \tag{13}$$

$$Recall = TP/(TP + FN) \tag{14}$$

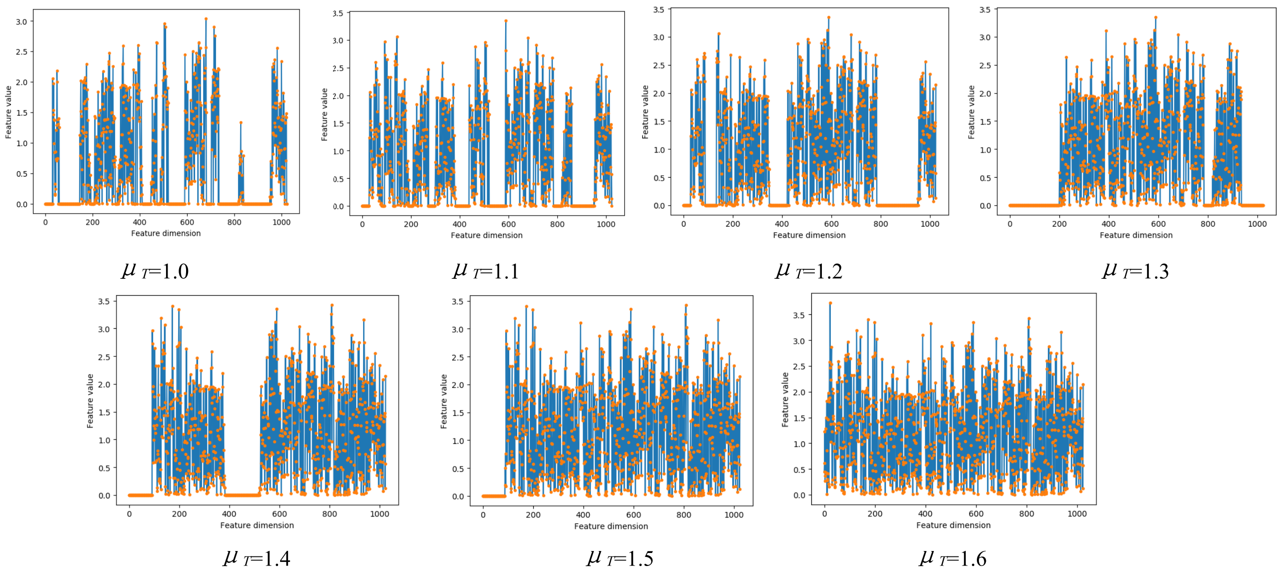
$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{15}$$

where *TP* is the number of true positive samples, *FN* is that of false negative samples, *FP* is that of false positive samples and *TN* is that of true negative samples.

B. RESUTLS ON DIFFERENT DATASETS

For the UCSDped1dataset and coal mining video dataset, the initialization of weights and biases variables are random values in the process of training VGG-LSTM networks. Moreover, we utilize the dropout function, parameter sharing [52] and data enhancement [53] methods to address the overfitting problem, and adopt the Adam algorithm to optimize the loss function. The change of weight variables and loss in the process of training for the UCSDped1 dataset and coal mining video dataset are shown in Figs. 5 and 6. From Fig. 5, we can see that the loss function is convergent, the different weights in CNN-LSTM networks change in the range of -0.2 to 0.2, and the different biases change in the range of -1 to 1. Moreover, according to Fig. 6, we can see that the loss function also is convergent, the different weights in CNN-LSTM networks change in the range of -0.2 to 0.2, and the different biases change in the range of -0.7 to 1.3. Therefore, there is no overfitting problem in the process of training and validation in our experiment.

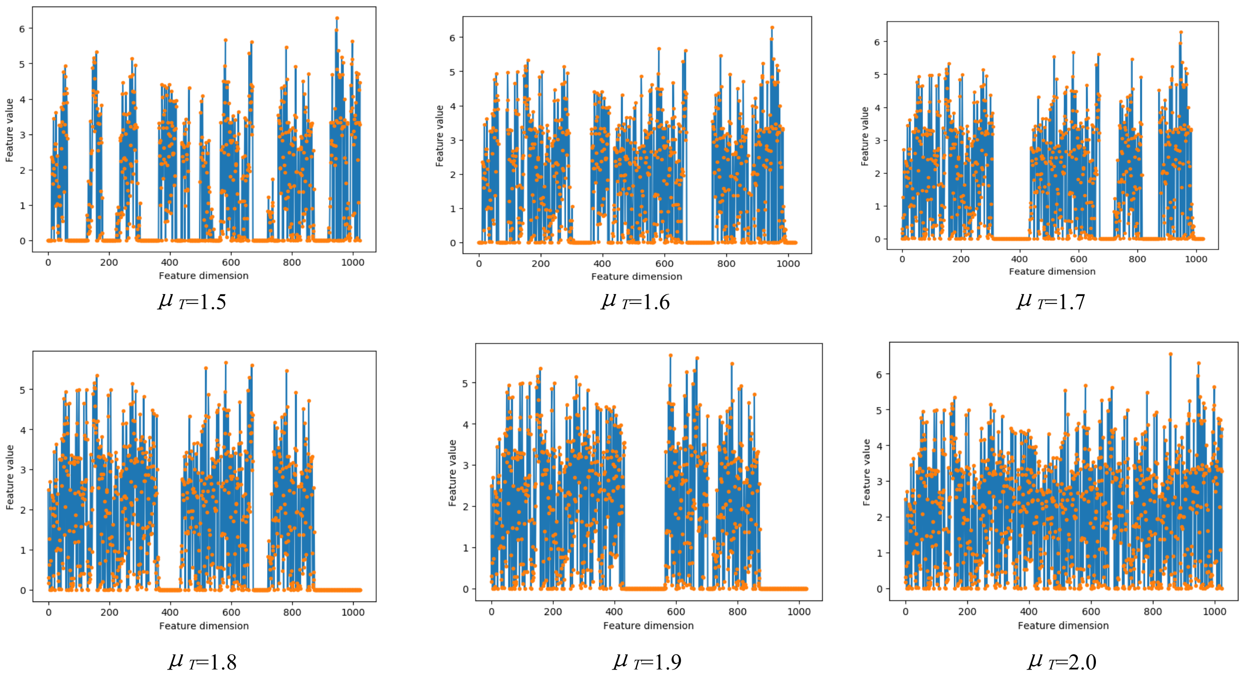
In our experiments, there are different feature expectation subgraphs that are constructed through different thresholds μ_T, as shown in Figs. 7 and 8. Through these figures, we can find that the number of eigenvalue points in the feature expectation subgraph increases with the increase of μ_T,



μ_T=1.0 μ_T=1.1 μ_T=1.2 μ_T=1.3

μ_T=1.4 μ_T=1.5 μ_T=1.6

FIGURE 7. The feature expectation subgraphs for different μ_T on UCSDped1dataset.



μ_T=1.5 μ_T=1.6 μ_T=1.7

μ_T=1.8 μ_T=1.9 μ_T=2.0

FIGURE 8. The feature expectation subgraphs for different μ_T on coal mining video dataset.

the topological structure of subgraphs in Fig. 7 changes until $\mu_T = 1.6$, and the topological structure of subgraphs in Fig. 8 also changes until $\mu_T = 2.0$. In addition, Figs. 9 and 10 show the situation that exists the less eigenvalue points, the higher accuracy. The accuracy is the highest when $\mu_T = 1.3$ in Fig.7 and $\mu_T = 2.0$ in Fig. 8. Less eigenvalue points in a feature expectation subgraph are not enough to represent features of video frame completely, and some feature expectation subgraphs or feature graphs may contain some eigenvalue points that correspond to disturbance factors, which will affect the accuracy of abnormal event detection.

To further study the performance of the proposed approach, we compare DF-ESCC with several state-of-the-art approaches. The results, as shown in Tables 1 and 2, demonstrate that the performance of hand-crafted features-based models is weaker than deep appearance features-based models, and our approach improves the performance effectively.

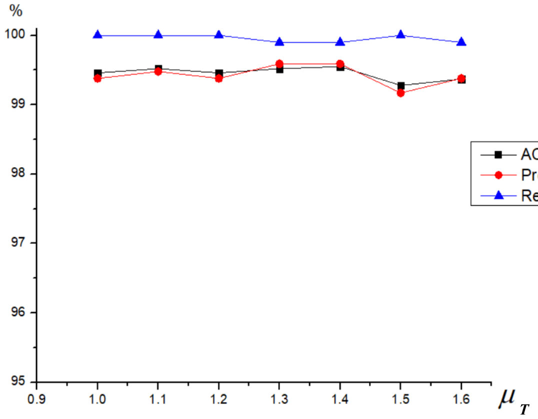


FIGURE 9. The performance of our method for different μ_T's on UCSD-ped1 dataset.

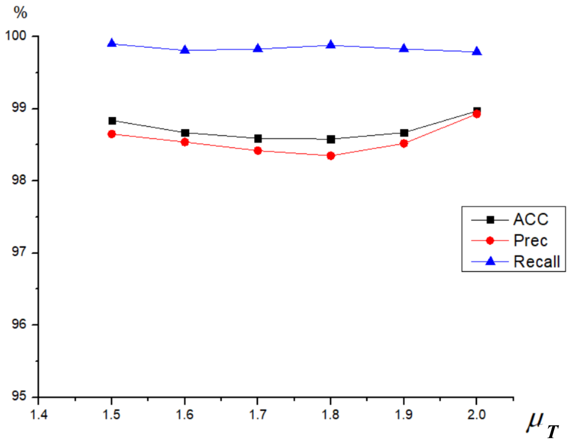


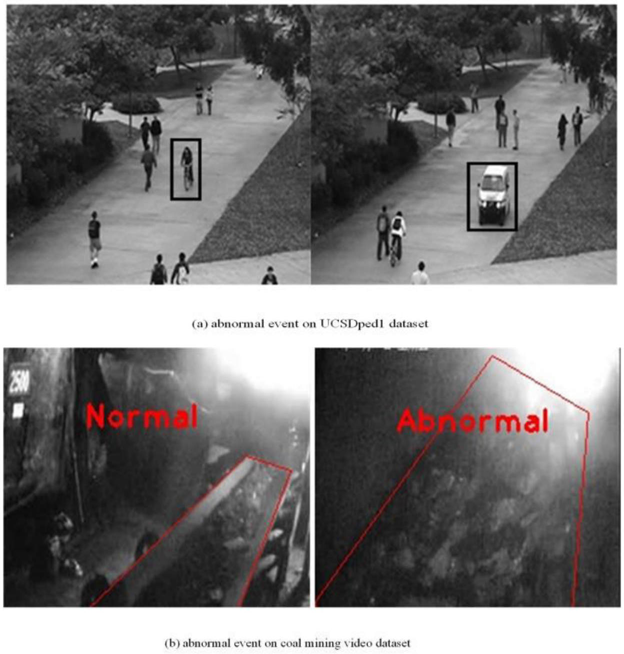
FIGURE 10. The performance of our method for different μ_T's on coal mining video dataset.

TABLE 1. Abnormal event detection on the ucsdped1 dataset.

Methods	Precision (%)	Recall (%)	Accuracy (%)
SURF+BoW	75.47	74.44	77.56
SIFT+BoW	74.25	79.63	76.23
HMM with optical flow	90.4	92.2	88.6
CNN-2D+LSTM	94.17	79.63	91.43
CNN-2D+LSTM+SVM	99.59	99.28	99.01
DF-ESCC	99.59	99.89	99.55

TABLE 2. abnormal event detection on the coal mining video dataset.

Methods	Precisi on (%)	Recall (%)	Accuracy (%)
SURF+BoW	65.43	72.69	73.13
SIFT+BoW	64.38	73.45	74.38
HMM with optical flow	77.6	90.6	83.1
CNN-2D+LSTM	98.76	99.64	92.30
CNN-2D+LSTM+SVM	98.34	99.19	98.03
DF-ESCC	98.93	99.79	98.97



(a) abnormal event on UCSDped1 dataset

(b) abnormal event on coal mining video dataset

FIGURE 11. The results of abnormal event detection on two datasets.

Some eigenvalue points correspond to the disturbances of light source in the coal mining video dataset and the dense crowd in UCSDped1 are filtered, which can reduce the influence of some disturbances on abnormal event detection. However, if the disturbances have a great influence on image features, the effect of maximum feature expectation graph will be better, such as in the case of $\mu_T = 2.0$ in Fig. 10. Finally, the results of abnormal event detection are shown in Fig. 11, where Fig. 11(a) shows the abnormal event that the car is on the pedestrian way, and Fig. 11(b) shows the abnormal event that coal accumulates on belt conveyor in the process of coal mining.

V. CONCLUSION

In this paper, we present an abnormal event detection hybrid modulation method via feature expectation subgraph calibrating classification (DF-ESCC) in video surveillance scenes. The proposed method based on feature extraction of VGG-16 and long short-term memory networks can extract the salient features accurately from surveillance videos. Moreover, some unexpected eigenvalues can be filtered by utilizing the constructed feature expectation subgraphs and mapping sparse vectors. Finally, the accuracy of abnormal event detection can be improved by using the classification of feature expectation subgraphs to calibrate the results of a single classifier. The experimental results on two challenging datasets indicate the effectiveness of DF-ESCC and show competitive performance with the existing approaches. In summary, the accuracy of abnormal event detection can be improved by utilizing internal sequential and topological relational characteristics of structured deep appearance features.

Although our approach can learn the effective discriminative features from CNN-LSTM networks, its performance still needs to improve in complex video surveillance scenes, and the graph kernel model also needs to improve. In the future, we plan to use inception networks and other graph kernel methods to further improve the performance of our method.

REFERENCES

[1] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 FPS in MATLAB," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2720–2727.

[2] A. D. Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 334–349.

[3] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *Int. J. Comput. Vis.*, vol. 90, no. 2, pp. 150–165, Nov. 2010.

[4] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao, "Interestingness prediction by robust learning to rank," in *Proc. 13th Eur. Conf. Comput. Vis.*, Zurich, Switzerland, Sep. 2014, pp. 488–503.

[5] M. J. Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2611–2618.

[6] W. Chu, H. Xue, C. Yao, and D. Cai, "Sparse coding guided spatiotemporal feature learning for abnormal event detection in large videos," *IEEE Trans. Multimedia*, vol. 21, no. 1, pp. 246–255, Jan. 2019.

[7] S. Smeureanu, R. T. Ionescu, M. Popescu, and A. Bogdan, "Deep appearance features for abnormal behavior detection in video," in *Proc. Int. Conf. Image Anal. Process.*, Catania, Italy, Sep. 2017, pp. 779–789.

[8] T. Fernando, S. Denman, S. Sridharan, and C. Fookes, "Soft + Hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection," *Neural Netw.*, vol. 108, pp. 466–478, Dec. 2018, doi: 10.1016/j.neunet.2018.09.002.

[9] Z. Fang, F. Fei, Y. Fang, L. Shu, and W. Wan, "Abnormal event detection based on saliency information," *Int. J. Multimedia Ubiquitous Eng.*, vol. 10, no. 9, pp. 339–352, Oct. 2015.

[10] R. Ye and X. Li, "Collective representation for abnormal event detection," *J. Comput. Sci. Technol.*, vol. 32, no. 3, pp. 470–479, May 2017.

[11] H. Chen, J. Gai, S. Zhang, C. Wang, C. Guo, X. Ye, and Y. Lu, "Abnormal event detection based on cosparse reconstruction," *J. Eng.*, vol. 2018, no. 5, pp. 254–256, May 2018.

[12] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream CNN," *IEEE Trans. Multimedia*, vol. 19, no. 7, pp. 1510–1520, Jul. 2017.

[13] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 1446–1453.

[14] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *Proc. CVPR*, Colorado, FL, USA, Jun. 2011, pp. 3313–3320.

[15] K.-W. Cheng, Y.-T. Chen, and W.-H. Fang, "Video anomaly detection and localization using hierarchical feature representation and Gaussian process regression," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 2909–2917.

[16] D. Singh and C. Krishna Mohan, "Graph formulation of video activities for abnormal activity recognition," *Pattern Recognit.*, vol. 65, pp. 265–272, May 2017, doi: 10.1016/j.patcog.2017.01.001.

[17] S. Bouindour, M. Hittawe, S. Mahfouz, and H. Snoussi, "Abnormal event detection using convolutional neural networks and 1-class SVM classifier," *IET Seminar Dig.*, vol. 2017, no. 5, pp. 1–6, Dec. 2017.

[18] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, San Francisco, CA, USA, Jun. 2010, pp. 1975–1981.

[19] O. Ye, Z. Li, and Y. Zhang, "Near-duplicate video cleansing method based on locality sensitive hashing and the sorted neighborhood method," in *Proc. 2nd EAI Int. Conf. Robot. Sensor Netw.*, Kitakyushu, Japan, Aug. 2018, pp. 129–139.

[20] F. Terroso-Saenz, M. Valdes-Vela, E. den Breejen, P. Hanckmann, R. Dekker, and A. F. Skarmeta-Gomez, "CEP-traj: An event-based solution to process trajectory data," *Inf. Syst.*, vol. 52, pp. 34–54, Aug. 2015, doi: 10.1016/j.is.2015.03.005.

[21] X. Song, L. Sun, J. Lei, D. Tao, G. Yuan, and M. Song, "Event-based large scale surveillance video summarization," *Neurocomputing*, vol. 187, pp. 66–74, Apr. 2016, doi: 10.1016/j.neucom.2015.07.131.

[22] S. Cosar, G. Donatiello, V. Bogorny, C. Garate, L. O. Alvares, and F. Bremond, "Toward abnormal trajectory and event detection in video surveillance," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 3, pp. 683–695, Mar. 2017.

[23] X. Wang, H. Song, and H. Cui, "Pedestrian abnormal event detection based on multi-feature fusion in traffic video," *Optik*, vol. 154, pp. 22–32, Feb. 2018, doi: 10.1016/j.ijleo.2017.09.104.

[24] S. A. Ahmed, D. P. Dogra, S. Kar, and P. P. Roy, "Surveillance scene representation and trajectory abnormality detection using aggregation of multiple concepts," *Expert Syst. Appl.*, vol. 101, pp. 43–55, Jul. 2018, doi: 10.1016/j.eswa.2018.02.013.

[25] P. Fu, H. Wang, K. Liu, X. Hu, and H. Zhang, "Finding abnormal vessel trajectories using feature learning," *IEEE Access*, vol. 5, pp. 7898–7909, 2017, doi: 10.1109/ACCESS.2017.2698208.

[26] S. Arif Ahmed, D. Prosad Dogra, S. Kar, and P. Pratim Roy, "Trajectory-based surveillance analysis: A survey," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 29, no. 7, pp. 1985–1997, Jul. 2019.

[27] Y. Cong, J. Yuan, and Y. Tang, "Video anomaly search in crowded scenes via spatio-temporal motion context," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 10, pp. 1590–1599, Oct. 2013.

[28] P. Mudjirahardjo, J. K. Tan, H. Kim, and S. Ishikawa, "Temporal analysis for fast motion detection in a crowd," *Artif. Life Robot.*, vol. 20, no. 1, pp. 56–63, Jan. 2015.

[29] J. Xu, S. Denman, C. Fookes, and S. Sridharan, "Detecting rare events using Kullback–Leibler divergence: A weakly supervised approach," *Expert Syst. Appl.*, vol. 54, pp. 13–28, Jul. 2016, doi: 10.1016/j.eswa.2016.01.035.

[30] P. M. Ashok Kumar and V. Vaidehi, "Anomalous event detection in traffic video based on sequential temporal patterns of spatial interval events," *KSII Trans. Internet Inf. Syst.*, vol. 9, no. 1, pp. 169–189, Jan. 2015.

[31] S. Liu, Y. Jin, Y. Tao, and X. Tang, "A novel representation for abnormal crowd motion detection," in *Proc. 7th Int. Conf. Intell. Sci. Big Data Eng.*, Dalian, China, Sep. 2017, pp. 239–248.

[32] A. B. Mabrouk and E. Zagrouba, "Spatio-temporal feature using optical flow based distribution for violence detection," *Pattern Recognit. Lett.*, vol. 92, pp. 62–67, Jun. 2017, doi: 10.1016/j.patrec.2017.04.015.

[33] T. Wang, M. Qiao, A. Zhu, Y. Niu, C. Li, and H. Snoussi, "Abnormal event detection via covariance matrix for optical flow based feature," *Multimedia Tools Appl.*, vol. 77, no. 13, pp. 17375–17395, Jul. 2018.

[34] T. Wang, M. Qiao, Y. Deng, Y. Zhou, H. Wang, Q. Lyu, and H. Snoussi, "Abnormal event detection based on analysis of movement information of video sequence," *Optik*, vol. 152, pp. 50–60, Jan. 2018, doi: 10.1016/j.ijleo.2017.07.064.

[35] F. Wang, Y. Ma, Y. Jin, Y. Jiang, and Y. Wang, "Discovering graphical visual features for abnormal semantic event detection," *Multimedia Tools Appl.*, vol. 77, no. 3, pp. 3245–3260, Feb. 2018.

[36] T. Gupta, V. Nunavath, and S. Roy, "CrowdVAS-net: A deep-CNN based framework to detect abnormal crowd-motion behavior in videos for predicting crowd disaster," in *Proc. IEEE Int. Conf. Syst., Man Cybern. (SMC)*, Bari, Italy, Oct. 2019, pp. 2877–2882.

[37] Y.-L. Hsueh, W.-N. Lie, and G.-Y. Guo, "Human behavior recognition from multiview videos," *Inf. Sci.*, vol. 517, pp. 275–296, May 2020, doi: 10.1016/j.ins.2020.01.002.

[38] V. Quan Nguyen, L. Van Ma, J.-y. Kim, K. Kim, and J. Kim, "Applications of anomaly detection using deep learning on time series data," in *Proc. IEEE 16th Int. Conf. Dependable, Autonomic Secure Comput., 16th Int. Conf. Pervas. Intell. Comput., 4th Int. Conf. Big Data Intell. Comput. Cyber Sci. Technol. Congr. (DASC/PiCom/DataCom/CyberSciTech)*, Aug. 2018, pp. 393–396.

[39] S. Latif, M. Usman, R. Rana, and J. Qadir, "Phonocardiographic sensing using deep learning for abnormal heartbeat detection," *IEEE Sensors J.*, vol. 18, no. 22, pp. 9393–9400, Nov. 2018.

[40] A. Santana, Y. Kawamura, K. Murakami, T. Iizaka, T. Matsui, and Y. Fukuyama, "Unsupervised fault detection in refrigeration showcase with single class data using autoencoders," *IEEJ Trans. Electron., Inf. Syst.*, vol. 139, no. 10, pp. 1191–1200, Oct. 2019.

[41] J. Fan, Q. Zhang, J. Zhu, M. Zhang, Z. Yang, and H. Cao, "Robust deep auto-encoding Gaussian process regression for unsupervised anomaly detection," *Neurocomputing*, vol. 376, pp. 180–190, Feb. 2020, doi: 10.1016/j.neucom.2019.09.078.

[42] Z. Fang, F. Fei, Y. Fang, C. Lee, N. Xiong, L. Shu, and S. Chen, "Abnormal event detection in crowded scenes based on deep learning," *Multimedia Tools Appl.*, vol. 75, no. 22, pp. 14617–14639, Nov. 2016.

[43] D. Arifoglu and A. Bouchachia, "Detection of abnormal behaviour for dementia sufferers using convolutional neural networks," *Artif. Intell. Med.*, vol. 94, pp. 88–95, Mar. 2019, doi: 10.1016/j.artmed.2019.01.005.

[44] H. Xia, T. Li, W. Liu, X. Zhong, and J. Yuan, "Abnormal event detection method in surveillance video based on temporal CNN and sparse optical flow," in *Proc. 5th Int. Conf. Comput. Data Eng. (ICCDE)*, Shanghai, China, 2019, pp. 90–94.

[45] J. Sun, J. Shao, and C. He, "Abnormal event detection for video surveillance using deep one-class learning," *Multimedia Tools Appl.*, vol. 78, no. 3, pp. 3633–3647, Feb. 2019.

[46] X. Zhang, R. Wang, and J. Ding, "Abnormal event detection by learning spatiotemporal features in videos," in *Proc. Chin. Conf. Image Graph. Technol.*, Beijing, China, Apr. 2018, pp. 421–431.

[47] K. Vignesh, G. Yadav, and A. Sethi, "Abnormal event detection on BMTT-PETS 2017 surveillance challenge," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Honolulu, HI, USA, Jul. 2017, pp. 2161–2168.

[48] J. T. Zhou, J. Du, H. Zhu, X. Peng, Y. Liu, and R. S. M. Goh, "AnomalyNet: An anomaly detection network for video surveillance," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 10, pp. 2537–2550, Oct. 2019.

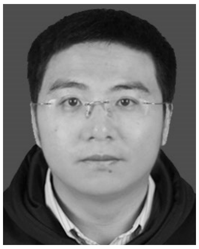
[49] H. Lin, Y. Hsueh, and W. Lie, "Convolutional recurrent neural networks for posture analysis in fall detection," *J. Inf. Sci. Eng.*, vol. 34, no. 3, pp. 577–591, May 2018.

[50] L. Wang, F. Zhou, Z. Li, W. Zuo, and H. Tan, "Abnormal event detection in videos using hybrid spatio-temporal autoencoder," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Athens, Greece, Oct. 2018, pp. 2276–2280.

[51] S. Yan, J. S. Smith, W. Lu, and B. Zhang, "Abnormal event detection from videos using a two-stream recurrent variational autoencoder," *IEEE Trans. Cognit. Develop. Syst.*, vol. 12, no. 1, pp. 30–42, Mar. 2020.

[52] J. Wang, W. Wang, S. Wei, Y. Zeng, and F. Luo, "Time series sequences classification with inception and LSTM module," in *Proc. IEEE Int. Conf. Integr. Circuits, Technol. Appl. (ICTA)*, Chengdu, China, Nov. 2019, pp. 51–55.

[53] L. Tian, Y. Zheng, and Q. Cui, "Research on data enhanced ancient pictogram recognition method based on convolutional neural network," in *Proc. 3rd High Perform. Comput. Cluster Technol. Conf. (HPCCT)*, Chengdu, China, 2019, pp. 210–214.



OU YE received the B.S. degree in computer science and engineer and the M.S. and Ph.D. degrees in computer software and theory and mechanical engineering from the Xi'an University of Technology, China, in 2007, 2010, and 2014, respectively.

He is currently an Associate Professor with the College of Computer Science and Technology, Xi'an University of Science and Technology. His current research interests include data cleansing, video retrieval, and image processing.

JUN DENG received the B.S. degree in mining engineering from the Xiangtan Mining College, in 1993, and the M.S. and Ph.D. degrees in mining engineering from the Department of Mining Engineering and Active College, Xi'an University of Technology, in 1996 and 2004, respectively.

He is currently a Professor with the College of Safety Science and Engineering, Xi'an University of Science and Technology, China. His current research interests include coal fire safety and public safety.

ZHENHUA YU received the B.S. and M.S. degrees from Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree from Xi'an Jiaotong University, Xi'an, in 2006.

He is currently a Professor with the College of Computer Science and Technology, Institute of Systems Security and Control, Xi'an University of Science and Technology, Xi'an. He has authored more than 20 technical articles for conferences and journals, and holds two invention patents. His research interests include cyber-physical systems and system security.

TAO LIU received the B.S. degree in computer science and technology from the Xi'an University of Science and Technology, in 2017. He is currently pursuing the M.S. degree with software engineering with the College of Computer Science and Technology, Xi'an University of Science and Technology. His research interests include video image/video abnormal detection and machine learning.

LIHONG DONG received the B.S. degree in computer education from the Xi'an Mining College, and the Ph.D. degree from the China University of Mining and Technology. She is currently a Professor with the College of Computer and Science and Technology, Xi'an University of Science and Technology, China. Her current research interests include software engineering and mining industry internet technology.