Sparse Coding Guided Spatiotemporal Feature Learning for Abnormal Event Detection in Large Videos

Wenqing Chu, Hongyang Xue, Chengwei Yao* and Deng Cai, *Member, IEEE*

***Abstract*—Abnormal event detection in large videos is an important task in research and industrial applications, which has attracted considerable attention in recent years. Existing methods usually solve this problem by extracting local features and then learning an outlier detection model on training videos. However, most previous approaches merely employ hand-crafted visual features, which is a clear disadvantage due to their limited representation capacity. In this paper, we present a novel unsupervised deep feature learning algorithm for the abnormal event detection problem. To exploit the spatiotemporal information of the inputs, we utilize the deep 3-dimensional convolutional network (C3D) to perform feature extraction. Then the key problem is how to train the C3D network without any category labels. Here, we employ the sparse coding results of the hand-crafted features generated from the inputs to guide the unsupervised feature learning. Specifically, we define a multi-level similarity relationship between these inputs according to the statistical information of the shared atoms. In the following, we introduce the quadruplet concept to model the multi-level similarity structure, which could be used to construct a generalized triplet loss for training the C3D network. Furthermore, the C3D network could be utilized to generate the features for sparse coding again, and this pipeline could be iterated for several times. By jointly optimizing between the sparse coding and the unsupervised feature learning, we can obtain robust and rich feature representations. Based on the learned representations, the sparse reconstruction error is applied to predicting the anomaly score of each testing input. Experiments on several publicly available video surveillance datasets in comparison with a number of existing works demonstrate that the proposed algorithm performs favorably against the state-of-the-art methods.**

***Index Terms*—Video analysis, unsupervised feature learning, sparse coding, anomaly detection.**

I. Introduction

In recent years, there have been growing interests in developing algorithms to automatically analyze large-scale multimedia data [1], [2], [3], [4], [5], [6], [7], [8], [9], [10], [11]. Among many video analysis tasks, detecting unusual events in video streams is of vital importance as it is related to other interesting topics in computer vision, such as visual saliency [12], dominant behavior detection [13] and interestingness prediction [14]. Unfortunately, abnormal event detection in video sequences is a difficult challenge due to the volatility of the definitions for both normality and abnormality.

This problem has drawn considerable attention from both academia and industry [15], [16], [17], [18], [19], [20], [21], [22]. A typical approach to address the anomaly detection problem is to extract local features and then tackle an outlier detection task where normal activities in video scenes are modeled, and unusual events are detected as they significantly diverge from the model. Among previous methods, trajectories are widely used features [23], [17], [24], due to their ability to describe the dynamic information of foreground objects. Besides, researchers also tackle this task by learning spatiotemporal activity patterns based on hand-crafted features, such as color, texture and optical flow [25], [26], [20], [21]. However, most hand-crafted features are designed under some priori knowledge that the intra-class variation is caused by some specific factor like illumination, scale or deformation. In case of complex video surveillance scene, these features do not generalize well.

In the past few years, deep neural networks based approaches have achieved remarkable progress in various computer vision applications, such as image classification [27], object detection [28] and activity recognition [29]. Although there are a number of deep learning methods [30], [31], [32] for dealing with video data. Most of them need large datasets of millions of labeled examples to learn rich and discriminative visual representations and can not be applied to abnormal event detection directly. As for the training data of abnormal event detection in video sequences, there are no class annotations, and all of them are viewed as normal events. One possible solution is the unsupervised deep learning approaches based on autoencoder networks which have been developed and investigated to address the abnormal event detection problem [33], [34], [35]. Nevertheless, the squared error based element-wise metric is not very suitable for visual data, as they do not model the properties of human visual perception [36]. In addition, these autoencoder networks [33], [34], [35] usually stack the neighbouring frames directly as input and perform convolution operation spatially [30] which can not exploit the spatiotemporal structure information well. Therefore, the absence of some rich and robust spatiotemporal features for describing the complex video scenes make anomaly detection still a very challenging task.

Recently, the deep 3-dimensional convolutional network (C3D) [30] has been used successfully in various applications such as action recognition [30], action localization [37], video

W. Chu, H. Xue and D. Cai are with State Key Lab of CAD&CG, College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. E-mail: wqchu16@gmail.com, hyxue@outlook.com, dengcai@gmail.com.

C. Yao is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China. E-mail: yaochw@zju.edu.cn.

*Corresponding author.

description [8], video recommendation [38], to name a few. The C3D is suitable for video data because the 3D convolution and 3D pooling operations can capture local spatiotemporal patterns better. However, it is difficult to apply this supervised model to abnormal event detection directly because there are no category labels for the training data of abnormal event detection and all of the video sequence data is viewed as normal events. There have been a large number of unsupervised deep feature learning works [39], [40], [41] for image data. The key insight of these works is to construct self-supervised signals by exploiting spatial information in the images or utilizing traditional unsupervised methods such as agglomerative clustering to obtain the similarity relationship between the inputs. This inspired us that we could employ the sparse coding method to obtain self-supervised signals for training the C3D network.

In this paper, we propose a novel Sparse Coding Guided Spatiotemporal Feature Learning (SCG-SF) framework to generate useful representations, which can be applied for modeling activity patterns for abnormal event detection effectively. Our algorithm consists of four parts, including C3D network, Dictionary learning, Constructing multiple level similarity trees and Triplet loss for a multi-level relationship. First, we employ the sparse coding results of the hand-crafted visual features to build multiple level similarity trees for these inputs. Following, we construct the quadruplet loss for a multi-level similarity structure, which could be used as supervised signals for training the C3D network. Furthermore, the C3D network could be used to generate the spatiotemporal features which could be used for sparse coding again. By alternative updating between the sparse coding and unsupervised feature learning, we can obtain robust and rich feature representations. After that, we apply the sparse reconstruction error to predicting the anomaly score of each testing input. The experimental results on several benchmark datasets indicate the effectiveness of our SCG-SF framework. In addition, the gains from the proposed SCG-SF are complementary to recent improvements in abnormal event detection which focus on outlier detection models.

Our contributions are summarized as follows:

- We propose an unsupervised deep feature learning framework to represent spatiotemporal information from videos. With the help of the sparse coding results, the learned representations will be more discriminative and representative.
- As far as we know, we are the first to introduce the powerful C3D network to extract spatiotemporal feature representations for abnormal event detection.
- The proposed method is validated on challenging anomaly detection datasets, and we obtain very competitive performance compared with the state-of-the-art methods.

The rest of this paper is organized as follows. In the next section, we present a brief review of the related works. Then we propose the Sparse Coding Guided Spatiotemporal Feature Learning framework in Section 3. The experimental results on three real-world datasets and comparison with a number of existing anomaly detection algorithms are presented in Section 4. Finally, we provide some concluding remarks in Section 5.

II. Related work

In this section, we briefly review previous methods considering abnormal event detection and related video analysis tasks. The most common approaches are based on two steps: extracting local features and learning an outlier detection model on training videos.

Among previous works, trajectories are widely used features for abnormal event detection [42], [17], [43], [44], [45], due to their ability of describing the dynamic information of foreground objects. By using accurate tracking algorithms, trajectory features can be extracted to model usual event patterns. In [42], trajectories which have similar spatial and velocity patterns are grouped and then used for building semantic scene models for detecting abnormal events. However, these tracking-based methods are not robust for complex scenes since they are composed of many components (blob detection, data association, tracking, ground plane trajectory extraction) and each of which may fail, leading to the failure of the whole anomaly detection system [46].

To overcome the drawback of trajectory based features, researchers address the problem by learning spatiotemporal activity patterns based on hand-crafted features extracted from low-level appearance and motion cues, such as color, texture and optical flow [47], [48], [49], [25], [50], [51], [26], [20], [52], [53], [21], [54]. In [50], a social force model based on optical flow features is proposed to represent regular event patterns and identify unusual events. Benezeth et al.[55] combine the co-occurrence statistics of spatiotemporal events with Markov Random Fields to perform outlier detection. In [56], a mixtures of dynamic textures model are introduced to employ appearance and motion features jointly. Cong et al. [20] design the spatiotemporal feature named multi-scale histograms of optical flow and then employ the sparse reconstruction error as a metric for identifying anomalous activities. In [57], a multi-task feature selection method is proposed to perform video action recognition with the Harris3D + HOG/HOF features. As an extension of sparse coding, the Graph Sparse Group Lasso model [58] is applied to the Local Binary Pattern features [59] of video shots and then the tags of the training videos are propagated to the test shots successfully according to the learned correlations. Furthermore, the Local Binary Pattern features and Scale-invariant feature transform features [60] from both images and video keyframes are extracted to learn a robust video indexing classifier [61]. In [62], the authors first learn a dictionary of features from local spatiotemporal cuboids and then measure the abnormality of an event according to the rarity of each feature in reconstructing all activities and the absolute coefficient used in reconstructing the current activity. However, most hand-crafted features are designed under some priori knowledge that the intra-class variation is caused by some specific factor like illumination, scale or deformation. In case of complex video surveillance scene, these features do not generalize well.

Recently, unsupervised deep learning approaches based on autoencoder networks have been developed and investigated

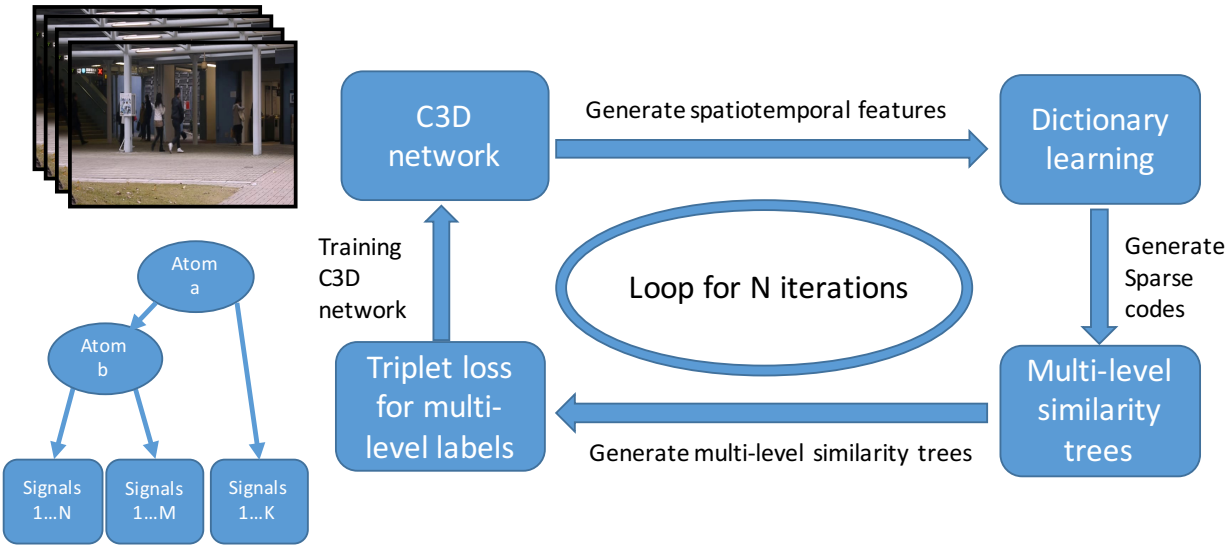


Fig. 1: The framework of the proposed algorithm. The proposed method is an iterative updating framework. For the first iteration, the C3D network has not been trained, and we can employ the sparse coding results of the hand-crafted features generated from the inputs to build multi-level similarity trees. And for the following iterations, the trained C3D network could be utilized to generate the spatiotemporal features for sparse coding again. Through jointly optimizing between the sparse coding and the unsupervised deep feature learning, we can obtain robust and rich feature representations.

to address the abnormal event detection problem [33], [34]. Hasan et al. [34] build a fully convolutional feedforward autoencoder and use reconstruction error for detecting abnormal events. In [33], stacked denoising autoencoders are proposed to learn rich and discriminative features containing both appearance and motion information. Then the learned feature representations are fed into multiple one-class SVM classifiers to predict the anomaly scores. Also, a cascade structure based on autoencoder networks [35] is proposed for fast anomaly detection. Furthermore, through reconstructing the temporal context in two directions instead of only the current frame, Zhu et al. [63] devise an unsupervised visual context learning module for video question answering. A multi rate visual recurrent model [64] also exploits the temporal context information for unsupervised video representation learning. For this autoencoder-based learning task, element-wise measures like the squared error are the default. Nevertheless, element-wise metrics are not very suitable for image data, as they do not model the properties of human visual perception (e.g., a small image translation might result in a large pixel-wise error whereas a human would barely notice the change [36]). Recently, the deep 3-dimensional convolutional network (C3D) has been used successfully in various applications [30], [37], [8] which inspired us to design C3D based feature representations for abnormal event detection. The C3D is suitable for spatiotemporal feature learning because the 3D convolution and 3D pooling operations in 3D ConvNet are able to model temporal information better. However, this model needs large datasets of millions of labeled examples to learn rich, high-performance visual representations and can not be applied to abnormal event detection directly. As for the training data of abnormal event detection in video sequences, there are no class annotations, and all of them are viewed as ordinary events.

In addition, some works consider an even more challenging setting, in which no additional training video streams are available [22], [65]. The approach proposed in [22] is to detect changes in a sequence of data from the video to see which frames are distinguishable from all the previous frames. And [65] labels a short-lasting event as abnormal if the amount of change from the immediately preceding event is substantially large. However, their performance is still not satisfactory for real applications.

III. The Proposed Algorithm

In this section, we describe how to develop a novel unsupervised deep feature for abnormal event detection in large videos. Recently, there has been a large number of works focusing on unsupervised deep feature learning [39], [66], [67], [40], [41] for image data. The key insight of these works is to construct self-supervised signals by exploiting spatial information in the images or utilizing traditional unsupervised methods such as agglomerative clustering to obtain the similarity relationship between the inputs. However, the agglomerative clustering is time-consuming, which limits its applicability to large-scale video data analysis. And autoencoder based approaches utilize the squared error as an element-wise measure which is simple but not very suitable for image data, as they do not model the properties of human visual perception [36]. In this paper, we try to combine the well-known sparse coding method with the unsupervised deep feature learning for video data. The core concept of sparse coding is that using an overcomplete dictionary that contains prototype atoms and then signals are

described by sparse linear combinations of these atoms [68]. The advantage of having an over-complete dictionary is that our basis vectors are better able to capture structures and patterns inherent in the input data. And the sparse representations could reflect the characteristic of the original inputs better. Here, we employ the sparse coding results of the hand-crafted features generated from the inputs to guide the unsupervised feature learning. Specifically, we define a multi-level similarity relationship between these inputs according to the statistical information of the shared atoms. In the following, the multi-level similarity relationship could be transformed into self-supervised signals for training the deep neural network.

The overview of our method is described in Figure 1. Our method consists of four parts, including C3D network, Dictionary learning, Constructing a multiple level similarity trees and Triplet loss for multi-level relationship. In the following, we will describe these parts separately, followed by the iterative updating framework.

A. C3D Network

The deep 3-dimensional convolutional network (C3D) has attracted considerable attention in recent years [30], [37], [8]. According to [30], 3D ConvNet is suitable for spatiotemporal feature learning because the 3D convolution and 3D pooling operations in 3D ConvNet are able to model temporal information better. Specifically, the 3D convolution and 3D pooling operations are performed spatiotemporally while in 2D ConvNets the 2D convolution and 2D pooling operations are done only spatially. Although some temporal stream network takes multiple frames as input, because of the 2D convolutions, after the first convolution layer, temporal information is collapsed completely. As far as we know, we are the first to introduce the powerful C3D network to extract spatiotemporal feature representations for abnormal event detection. We will describe that how to obtain the self-supervised signals in the next section. In this section, we focus on the design of the network architecture for this abnormal event detection problem. The whole architecture is demonstrated in Figure 2. According to the findings in 3D ConvNet [30], small receptive fields of $3 \times 3 \times 3$ convolution kernels with deeper architectures yield best results. Hence, for our architecture, we fix the spatiotemporal receptive field to $3 \times 3 \times 3$ with stride 1 in both spatial and temporal dimensions. Suppose the video clips are with a size of $c \times l \times h \times w$ where c is the number of channels, l is the frame number of the video clips, h and w are the height and width of the frame, respectively. In our experiments, we set the number of frames to 8. Therefore, after the 5th convolution layers, we will obtain a $128 \times 1 \times h/16 \times w/16$ tensor. For every location in the feature maps denoted as $\mathbf{y}_i \in \mathbb{R}^{128}$, it will describe a video cuboid of size $c \times 8 \times 16 \times 16$. Since we will integrate a metric learning based loss function in the following, we apply L2 normalization for each spatiotemporal feature $\mathbf{y}_i$.

B. Dictionary Learning

As mentioned before, we will utilize the sparse coding results to guide the unsupervised deep feature learning. In this section, we briefly describe how to perform dictionary learning for the generated spatiotemporal features and then obtain the sparse coding results. Suppose we have a lot of features $\mathbf{S} = \{\mathbf{y}_i\}_{i=1}^n$ and $\mathbf{y}_i \in \mathbb{R}^m$. And we want to obtain an overcomplete dictionary $\mathbf{D} \in \mathbb{R}^{m \times d}$, where $m < d$, so that the features $\mathbf{y}_i$ could be described by sparse linear combinations $\mathbf{x}_i$ of these atoms. The formulation is as below:

$$\underset{\mathbf{D},\mathbf{x}}{\operatorname{argmin}} \sum_{i=1}^{n} ||\mathbf{y}_i - \mathbf{D}\mathbf{x}_i||_2^2 + \lambda ||\mathbf{x}_i||_0 \tag{1}$$

Extraction of the sparsest representation is a hard problem and fortunately, it has been extensively investigated in the past few years. Considering the efficiency and accuracy, we utilize the algorithm KSVD proposed in [68] to complete this task. With the learned dictionary, we can employ the OMP algorithm to find a representation $\mathbf{x}_i$ with the fixed coefficient number for each spatiotemporal feature. The dictionary size d in the proposed method is a hyperparameter. We perform cross-validation with the learned representations and set the dictionary size to 500 according to the performance on the validation dataset.

C. Constructing multiple level similarity trees

With the sparse representations $\mathbf{x}_i$, we can capture structures and patterns inherent in the input data. The atoms indicate the basis vectors in the over-complete dictionary $\mathbf{D}$ and the signals indicate the features $\mathbf{y}_i$. Each atom could be regarded as an attribute or a feature. If two inputs share some specific basis vector and the corresponding coefficients are close, it's obvious that they could be similar to each other in terms of appearance and semantic. That inspired us to define a multi-level similarity relationship between these inputs according to the statistical information of the shared atoms. Figure 3 illustrates an example of two-level similarity trees. For some atom a, we first divide all of the signals into two parts determined by whether these signals are dominated by it (have large absolute coefficients). The formulation is as below:

$$\begin{cases} Left\ part = \left\{ \mathbf{y}_i | if\ ||\mathbf{x}_i^a||_1 \geq \dfrac{2}{n} \displaystyle\sum_{i=1}^{n} ||\mathbf{x}_i^a||_1 \right\}, \\ Right\ part = \left\{ \mathbf{y}_i | if\ ||\mathbf{x}_i^a||_1 \leq \dfrac{1}{n} \displaystyle\sum_{i=1}^{n} ||\mathbf{x}_i^a||_1 \right\}. \end{cases} \tag{2}$$

In the following, we find another atom b and divide the first part signals again like the Figure 3. In this fashion, we could build a tree-like hierarchy based on the statistical information of the shared atoms. The hierarchy can contain multiple levels. We also noticed that in [69], the authors utilized a tree-like hierarchy based on domain knowledge on the fine-grained car dataset as well. The difference is that they have fine-grained annotations for the training datasets and we do not have any category labels. Therefore, we employ the sparse coding results to generate the multiple level similarity trees. This operation could be done in a batch way. That means we need deal with hundreds of signals at each time which is very fast.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2846411, IEEE Transactions on Multimedia

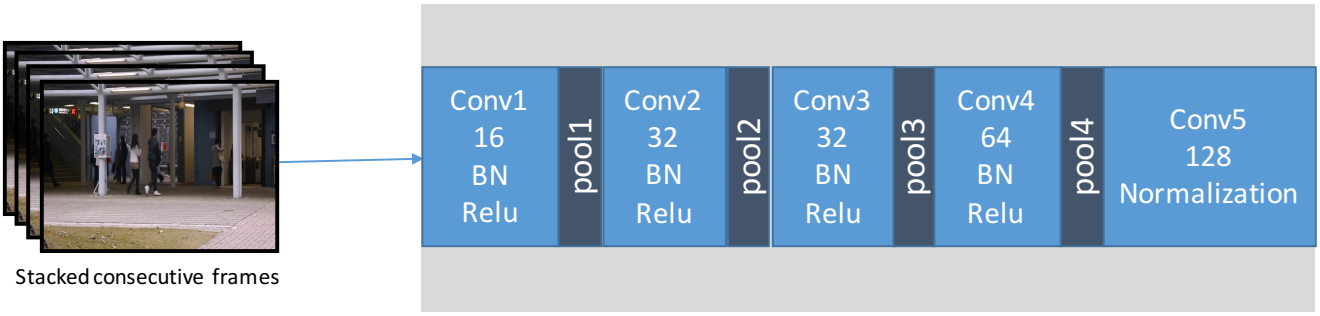


Fig. 2: The C3D network architecture. The C3D net has 4 3D convolution, 4 3D max-pooling, and 1 2D convolution layers, followed by a normalization layer. All 3D convolution kernels are $3\times3\times3$ with stride 1 in both spatial and temporal dimensions. The numbers of filters are denoted in each box. The 3D pooling layers are denoted from pool1 to pool4. All pooling kernels are $2\times2\times2$, except for pool1 is $1\times2\times2$. The last 2D convolution layer kernels are 3×3 with stride 1 in spatial dimensions.

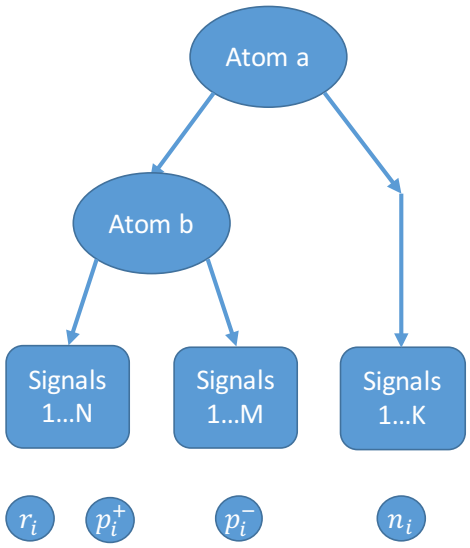


Fig. 3: A sample of multiple level similarity trees. For some atom a, we first divide all of the signals into two parts determined by whether these signals are dominated by it (have large absolute coefficients). In the following, we find another atom b and divide the first part signals again.

D. Triplet loss for multi-level relationship

After we obtain the multiple level similarity trees, we can construct a metric learning based loss function for training the C3D network. In our problem, we mainly focus on the scenario of a two-level similarity hierarchy like Figure 3. To model this hierarchy of similarity trees, we propose to generalize the concept of triplet following [69]. Specifically, quadruplet is introduced to model the two-level structure. Each quadruplet, (r_i, p_i^+, p_i^-, n_i), consists of four signals. Similar to the triplet, p_i^+ denotes the signals at the same leaf node in the second level as the reference r_i. The main difference is that in quadruplet, all negative samples are classified into two sub-categories: the more similar one p_i^- that shares the same parent node at the second level with r_i, and the more different one n_i sampled from the different branch at the first level. Given a quadruplet, this hierarchical relation among the four signals can be described in two inequalities,

$$D(r_i, p_i^+) + m_1 < D(r_i, p_i^-) + m_2 < D(r_i, n_i)$$

where the two hyper-parameters, m_1 and m_2, satisfying $m_1 > m_2 > 0$, control the distance margins across the two levels. In this paper, the D indicates the Euclidean distance. Compared to the triplet, quadruplet is able to model much richer similarity structures between different levels. As a result, the learned spatiotemporal feature representations can preserve the geometric structure of the input data better.

E. Iterative updating framework

In this section, we present the proposed iterative updating framework. For the first iteration, we can employ the sparse coding results of the hand-crafted features generated from the inputs to build multi-level similarity trees. We used 3d gradient features in this work. Then the multi-level similarity trees could be used to construct a generalized triplet loss for training the C3D network. And for the following iterations, the trained C3D network could be utilized to generate the spatiotemporal features for sparse coding again. Through jointly optimizing between the sparse coding and the unsupervised deep feature learning, we can obtain robust and rich feature representations. The whole procedure of the proposed sparse coding guided spatiotemporal feature learning method is summarized in Algorithm 1.

Based on the learned representations, the sparse reconstruction error is used to predict the anomaly scores of each testing input. In our work, we employ the sparse combination learning method proposed in [21] which is very efficient. The method in [21] has publicly available code. However, their code for the part of sparse combination learning is in pcode for the purpose of protecting its proprietary source code, leading to it is not proper for other features as this part contains a lot of hyper-parameters. Therefore, we have implemented this part by ourselves.

Algorithm 1 Sparse Coding Guided Spatiotemporal Feature Learning

Input: iteration number T, batch size b and training examples $\{\chi_n\}_{n=1}^N$,
Output: Parameters of C3D network,
Compute training examples' hand-crafted visual features $\{\mathbf{y}_n\}_{n=1}^N$ according to training examples $\{\chi_n\}_{n=1}^N$
for $t = 1$ to T **do**
  for $r = 1$ to N/b **do**
    Estimate a dictionary $\mathbf{D}$ on $\{\mathbf{y}_n\}_{n=1}^N$ with KSVD
    Obtain sparse representations $\{\mathbf{x}_n\}_{n=1}^N$ with OMP algorithm
    Build the multi-level similarity trees
    Construct triplet loss
    Update the C3D network parameters according to the triplet loss
  end for
  Update training examples' visual features $\{y_n\}_{n=1}^N$ according to training examples $\{\chi_n\}_{n=1}^N$
end for
Return: the trained C3D network

IV. Experiments

We conduct extensive experiments on three widely used abnormal event detection datasets to evaluate the performance of the proposed SCG-SF for unsupervised deep feature learning and compare them with several state-of-the-art methods. The following describes the details of the experiments and results.

A. Dataset

The Avenue, Subway and UCSD datasets are among the most commonly used benchmarks for abnormal event detection in videos. Following [22], we also mainly focus on the Avenue Dataset because it is more challenging than staged datasets (such as the UMN abnormal event detection dataset) and is more recent with more specific labeling than others. The dataset is also valuable because the method in [21], [22], [65] has publicly available code and results, so we could compare with the same implementation as a recent standard in anomaly detection.

1) Avenue dataset: The Avenue dataset [21] contains 16 training and 21 test videos. Each sequence is about 2 minutes long. In total, there are 15328 frames in the training set and 15324 frames in the test set. Each frame is 640×360 pixels. There are 14 unusual events including running, throwing objects, and loitering. Locations of anomalies are annotated in ground truth pixel-level masks for each frame in the testing videos.

2) Subway dataset: The Subway dataset [47] includes two videos: entrance gate (1 hour 36 minutes in duration consisting of 144,249 frames) and exit gate (43 minutes in duration consisting of 64,900 frames) with a 512×384 pixel resolution. Normal behaviors include people entering and exiting the station while unusual events include people moving in the wrong direction, such as exiting the entrance and entering the exit or avoiding payment.

3) UCSD dataset: The UCSD Pedestrian dataset [56] has two different scenes, called Ped1 and Ped2. The first scene Ped1 provides 34 training clips and 36 testing clips. Each clip has around 200 frames with a 238×158 pixel resolution. The second scene Ped2 contains 16 training clips and 12 testing clips. Each clip has a resolution of 360×240.

B. Evaluation Criteria

Following [20], the frame level criterion is used for evaluation of abnormal event detection accuracy based on true-positive rates (TPR) and false-positive rates (FPR). An algorithm determines the frames that contain abnormal events. The result is compared to the frame-level ground truth annotation of each frame, and the number of true and false positive frames are calculated. The two measures are combined into a receiver operating characteristic (ROC) curve of TPR versus FPR:

$$TPR = \frac{number\ of\ true\ positive\ frames}{number\ of\ positive\ frames}$$

$$FPR = \frac{number\ of\ false\ positive\ frames}{number\ of\ negative\ frames}$$

The TPR is calculated as the ratio between the number of true positive frames and number of positive frames, and the FPR is calculated as the ratio between the number of false positive frames and number of negative frames. TPR and FPR are calculated for different threshold values in the range from minimum to maximum anomaly score. Then, ROC curve is drawn as the TPR vs. FPR. Finally, the performance is summarized using the area under the curve (AUC) for frame level criterion. A high AUC value indicates better performance. Although many works [20], [62], [21] include the Equal Error Rate (EER) as the evaluation metric, we agree with [22], [65] that metrics such as the EER can be misleading in a realistic anomaly detection setting, in which abnormal events are expected to be very rare. Thus, we do not use the EER in our evaluation on the Avenue dataset. In addition, we also compute the pixel-level AUC referring to the code in [1] provided by [22]. For the Subway dataset, we follow the event level evaluation in [26]. Once the algorithm detects at least one frame in the annotated range, the detection is counted as correct. On the other hand, the false alarm is also measured in the same way: at least one frame is fired outside the annotated range, then it is counted as the false alarm.

C. Performance Evaluation

Our SCG-SF model has two essential parameters: the dictionary size of the dictionary learning part and the iteration number of the alternative updating. Here, we first conduct experiments on the Avenue dataset to evaluate the property of our method. Figure 4 shows how the dictionary size affects the performance of the proposed SCG-SF. As we can see that when the dictionary size becomes larger, our method achieves better frame AUC at first and then the performance decreases. And our method achieves the best performance when the dictionary size is set to 500. It demonstrates that there exists a balance

[1] https://alliedel.github.io/anomalydetection/

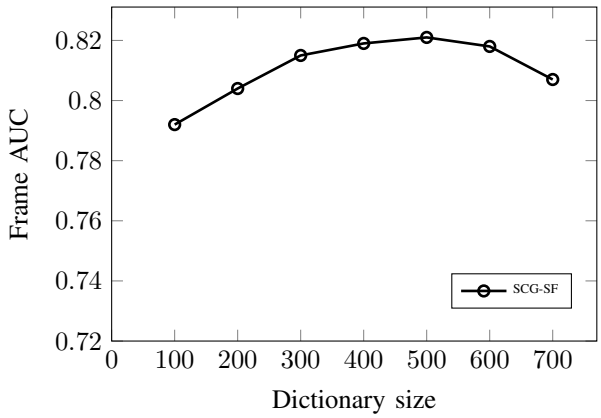


Fig. 4: Frame AUC on the Avenue dataset with different dictionary size. From the results, we can observe that SCG-SF achieves better performance when the dictionary size is 500.

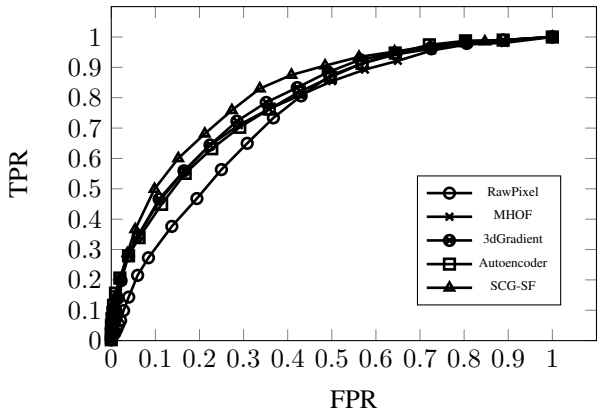


Fig. 5: Frame AUC on the Avenue dataset with different types of spatiotemporal features. As we can see, the performance of the RawPixel is the worst. And the performance of hand-craft features such as MHOF and 3dGradient are similar with the Autoencoder based features. And the proposed SCG-SF achieves the best performance among all of the spatiotemporal features.

between the reconstruction error and the generalization ability. Hence we use 500 as the dictionary size for performing the dictionary learning part in the following experiments. For the iteration number of the alternative updating, we find that SCG-SF generally achieves better performance when the iteration number goes up. However, we also notice that too many iterations would ruin the abnormal event detection performance. From the experimental results, we guess that after several iterations, the sparse coding maybe too aggressive and could not provide accurate supervised information. Therefore, in the experiments, we use the C3D network after three iterations to generate the evaluation features.

In addition, we also want to investigate the performance of the proposed SCG-SF compared with the following popular spatiotemporal features, including RawPixel, MHOF [20], 3dGradient [25], Autoencoder [70]. The learned representations will be used as input for the sparse combination learning method (SCL) [21]. The RawPixel features are extracted from a $10 \times 10 \times 5$ (rows $\times$ columns $\times$ frames) spatiotemporal subunits in the video. For the MHOF features, we employ the type C spatiotemporal basis in [20] which could corporate both local spatial and temporal information. The 3dGradient features are computed with the provided code in [22]. And for the autoencoder network, the encoder part is defined as $500 \rightarrow 512 \rightarrow 256 \rightarrow 128$, and the decoder part is a symmetric structure. After PCA and normalization, all of the features are represented by a 128-dimensional vector.

Figure 5 shows the frame-level ROC curves for different types of spatiotemporal features on the Avenue dataset. As we can see, the performance of the RawPixel is the worst. And the performance of hand-craft features such as MHOF and 3dGradient are similar with the Autoencoder based features. And the proposed SCG-SF achieves the best performance among all of the spatiotemporal features. Besides, we show the frame-level AUC and pixel-level AUC on the Avenue dataset in Table I. Furthermore, we want to investigate the performance of the generic pre-trained C3D features and thus we use the pre-trained model on UCF101 [71] provided in [2] to extract spatiotemporal features denoted by UCF101. This experimental result also demonstrated that our SCG-SF is very effective compared with other hand-crafted features or autoencoder based features under the same outlier detection method [21]. The frame-level AUC for UCF101 on the avenue dataset is around 80.1 which is similar to those of the hand-crafted features. We believe that there may be a huge gap between the source domain and the target domain.

To further study the performance of our approach, we compare SCG-SF with the following popular abnormal event detection methods, including SCL [21], Conv-AE [34], Spatiotemporal Autoencoder[72], Discriminative learning (DL) [22], Unmasking [65] and Stream Restricted Boltzmann machine (SRBM) [73]. Results are shown in Table II for our method with comparisons to related work. The experimental results of these compared methods are extracted from their paper. The 'X's in Table II mean that the pixel-level AUC is not provided. Despite the fact that we use no other regularization nor any hyper-parameter search, the proposed algorithm performs favorably against the state-of-the-art methods in both frame-level and pixel-level AUC. In addition, Figure 6 presents some examples of our detection results on Avenue dataset. The left figures are the original images. And in the right figures, we show the detection results by adding masks where blue indicates true positive and red indicates false positive. We use the left figures to show the appearance of the images without masks. We can see that the abnormal events in the top two rows can be detected correctly. However, our method generates a false positive detection around the security guard, and we think the reason is that the security guard seldom appears in the training videos.

In addition, we also evaluate the SCG-SF's performance on the Subway dataset and the UCSD dataset. We compare our method with many state-of-the-art abnormal event detection methods consisting of Incremental Coding Length (ICL) [62], Spatio-temporal Compositions (STC) [74], MPPCA [49],

[2]https://github.com/DavideA/c3d-pytorch

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2018.2846411, IEEE Transactions on Multimedia

TABLE I: Anomaly detection performance on the Avenue dataset for different features.The proposed SCG-SF performs favorably against other spatiotemporal features in both frame-level and pixel-level AUC.

Methods	Frame AUC	Pixel AUC
RawPixel	74.6	89.7
3dGradient	79.5	92.5
Autoencoder	78.7	91.8
MHOF	78.2	90.4
Ours	**82.1**	**93.7**

TABLE II: Anomaly detection performance on the Avenue dataset. The proposed algorithm performs favorably against the state-of-the-art methods in both frame-level and pixel-level AUC.

Methods	Frame AUC	Pixel AUC
SCL	80.9	92.9
Conv-AE	70.2	X
SRBM	78.76	X
Spatiotemporal Autoencoder	80.3	X
Discriminative learning	78.3	91.0
Unmasking	80.6	93.0
Ours	**82.1**	**93.7**

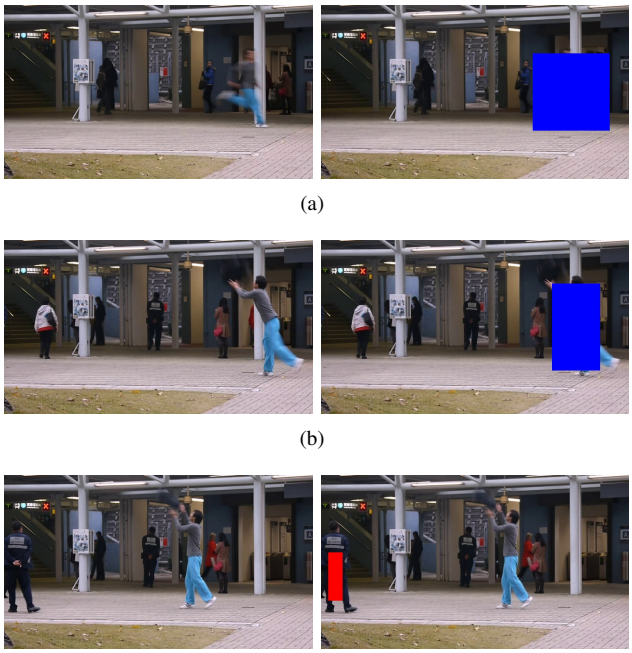
Social Force model (SF) [50], Local Statistical Aggregates (LSA) [53], Mixture of Dynamic Textures (MDT) [56], SCL [21], Dynamic SC [26], Sparse reconstruction cost (SRC) [20], Conv-AE [34], AMDN [33] and Deep Cascade [35]. The experimental results of these compared methods are extracted from their paper. The 'X's in Table IV and Table III mean that the results are not provided. Results for the Subway dataset are shown in Table IV for our method with comparisons to related work. As we can see, SCG-SF achieves comparable performance with respect to existing methods. And we show the results for the UCSD dataset in Table III in terms of EER (Equal Error Rate) and AUC (Area Under ROC). It demonstrates that our method is comparable to several state-of-the-art approaches even though we use simple sparse reconstruction error model instead of other complex outlier detection approaches. We notice that combining optical flow based motion features and appearance features like AMDN [33] could perform better on the UCSD dataset since most of the anomaly events have strong motion changes. Also, different outlier detection models have a notable influence on the accuracy performance for the abnormal event detection task.

TABLE III: Quantitative comparison between different methods on the UCSD dataset in terms of EER (Equal Error Rate) and AUC (Area Under ROC).

Methods	UCSDped1 AUC/EER	UCSDped2 AUC/EER
ICL	X/19.8	X/22.3
LSA	92.7/19.8	X/X
MDT	81.8/25.0	82.9/25.0
MPPCA	59.0/40.0	69.3/30.0
Conv-AE	81.0/27.9	90.0/21.7
AMDN	92.1/16.0	90.8/17.0
SCL	91.8/15.0	X/X
Deep Cascade	X/9.1	X/8.2
Ours	90.9/16.2	90.2/17.3

TABLE IV: Quantitative comparison between different methods on Subway dataset.

Methods	Dataset	Abnormal events	False alarm
ICL	Entrance Exit	60/66 19/19	5 2
STC	Entrance Exit	61/66 19/19	4 2
MPPCA	Entrance Exit	57/66 19/19	6 3
Dynamic SC	Entrance Exit	60/66 19/19	5 2
SRC	Entrance Exit	27/31 9/9	4 0
SCL	Entrance Exit	57/66 19/19	4 2
Ours	Entrance Exit	60/66 19/19	5 2



(a) (b) (c)

Fig. 6: True positive (top two rows) versus false positive (bottom row) detections of the proposed method based on the unsupervised deep learning features SCG-SF. Examples are selected from the Avenue dataset.

V. Conclusion

In this paper, we presented a novel unsupervised spatiotemporal feature learning approach for video anomaly detection. The proposed method is based on jointly optimizing the sparse coding and the unsupervised feature learning for video data. Then the learned representations will be utilized to feed into the sparse reconstruction error based model for outlier detection. The experimental results on two challenging datasets indicate the effectiveness of the proposed SCG-SF features and show competitive performance with respect to existing methods. Although our approach learns effective discriminative features, its performance is still far away from satisfaction while the supervised deep neural networks have achieved great progress. In the future, we plan to investigate alternative approaches for obtaining self-supervised signals such as the temporal order information in the training videos. In addition, we will try to extend our framework to other video analysis tasks where the labeling effort is also expensive.

References

[1] Y. Xie, H. Chang, Z. Li, L. Liang, X. Chen, and D. Zhao, "A unified framework for locating and recognizing human actions," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 25–32.

[2] S. H. Khatoonabadi and I. V. Bajic, "Video object tracking in the compressed domain using spatio-temporal markov random fields," *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 300–313, 2013.

[3] Y.-G. Jiang, Q. Dai, T. Mei, Y. Rui, and S.-F. Chang, "Super fast event recognition in internet videos," *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1174–1186, 2015.

[4] D. Coppi, S. Calderara, and R. Cucchiara, "Transductive people tracking in unconstrained surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 4, pp. 762–775, 2016.

[5] Y. Li, R. Wang, Z. Cui, S. Shan, and X. Chen, "Spatial pyramid covariance-based compact video code for robust face retrieval in tv-series," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5905–5919, 2016.

[6] K. R. Jerripothula, J. Cai, and J. Yuan, "Cats: co-saliency activated tracklet selection for video co-localization," in *European Conference on Computer Vision*. Springer, 2016, pp. 187–202.

[7] A. Montes, A. Salvador, and X. Giro-i Nieto, "Temporal activity detection in untrimmed videos with recurrent neural networks," *arXiv preprint arXiv:1608.08128*, 2016.

[8] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4594–4602.

[9] Y. Yan, F. Nie, W. Li, C. Gao, Y. Yang, and D. Xu, "Image classification by cross-media active learning with privileged information," *IEEE Transactions on Multimedia*, vol. 18, no. 12, pp. 2494–2502, 2016.

[10] J. Yuan, B. Ni, X. Yang, and A. A. Kassim, "Temporal action localization with pyramid of score distribution features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3093–3102.

[11] L. Sun, X. Wang, Z. Wang, H. Zhao, and W. Zhu, "Social-aware video recommendation for online social groups," *IEEE Transactions on Multimedia*, vol. 19, no. 3, pp. 609–618, 2017.

[12] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency estimation in video," *International Journal of Computer Vision*, vol. 90, no. 2, pp. 150–165, 2010.

[13] M. Javan Roshtkhari and M. D. Levine, "Online dominant and anomalous behavior detection in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2611–2618.

[14] Y. Fu, T. M. Hospedales, T. Xiang, S. Gong, and Y. Yao, "Interestingness prediction by robust learning to rank," in *European Conference on Computer Vision*. Springer, 2014, pp. 488–503.

[15] R. Hamid, A. Johnson, S. Batta, A. Bobick, C. Isbell, and G. Coleman, "Detection and explanation of anomalous activities: Representing activities as bags of event n-grams," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 1031–1038.

[16] D. Zhang, D. Gatica-Perez, S. Bengio, and I. McCowan, "Semi-supervised adapted hmms for unusual event detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1. IEEE, 2005, pp. 611–618.

[17] C. Piciarelli, C. Micheloni, and G. L. Foresti, "Trajectory-based anomalous event detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 11, pp. 1544–1554, 2008.

[18] A. Basharat, A. Gritai, and M. Shah, "Learning object motion patterns for anomaly detection and improved object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[19] F. Jiang, Y. Wu, and A. K. Katsaggelos, "A dynamic hierarchical clustering method for trajectory-based unusual video event detection," *IEEE Transactions on Image Processing*, vol. 18, no. 4, pp. 907–913, 2009.

[20] Y. Cong, J. Yuan, and J. Liu, "Sparse reconstruction cost for abnormal event detection," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3449–3456.

[21] C. Lu, J. Shi, and J. Jia, "Abnormal event detection at 150 fps in matlab," in *IEEE International Conference on Computer Vision*, 2013, pp. 2720–2727.

[22] A. Del Giorno, J. A. Bagnell, and M. Hebert, "A discriminative framework for anomaly detection in large videos," in *European Conference on Computer Vision*. Springer, 2016, pp. 334–349.

[23] C. Stauffer and W. E. L. Grimson, "Learning patterns of activity using real-time tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 747–757, 2000.

[24] Y. Shi, Y. Tian, Y. Wang, and T. Huang, "Sequential deep trajectory descriptor for action recognition with three-stream cnn," *IEEE Transactions on Multimedia*, 2017.

[25] L. Kratz and K. Nishino, "Anomaly detection in extremely crowded scenes using spatio-temporal motion pattern models," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1446–1453.

[26] B. Zhao, L. Fei-Fei, and E. P. Xing, "Online detection of unusual events in videos via dynamic sparse coding," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3313–3320.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Neural Information Processing Systems*, 2012, pp. 1097–1105.

[28] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[29] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1725–1732.

[30] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

[31] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 20–36.

[32] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Neural Information Processing Systems*, 2014, pp. 568–576.

[33] D. Xu, E. Ricci, Y. Yan, J. Song, and N. Sebe, "Learning deep representations of appearance and motion for anomalous event detection," in *British Machine Vision Conference*, 2015.

[34] M. Hasan, J. Choi, J. Neumann, A. K. Roy-Chowdhury, and L. S. Davis, "Learning temporal regularity in video sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 733–742.

[35] M. Sabokrou, M. Fayyaz, M. Fathy, and R. Klette, "Deep-cascade: cascading 3d deep neural networks for fast anomaly detection and localization in crowded scenes," *IEEE Transactions on Image Processing*, vol. 26, no. 4, pp. 1992–2004, 2017.

[36] A. B. L. Larsen, S. K. Sønderby, H. Larochelle, and O. Winther, "Autoencoding beyond pixels using a learned similarity metric," in *International Conference on Machine Learning*, 2016, pp. 1558–1566.

[37] Z. Shou, D. Wang, and S.-F. Chang, "Temporal action localization in untrimmed videos via multi-stage cnns," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1049–1058.

[38] T. Han, H. Yao, C. Xu, X. Sun, Y. Zhang, and J. J. Corso, "Dancelets mining for video recommendation based on dance styles," *IEEE Transactions on Multimedia*, vol. 19, no. 4, pp. 712–724, 2017.

[39] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *IEEE International Conference on Computer Vision*, 2015, pp. 1422–1430.

[40] J. Yang, D. Parikh, and D. Batra, "Joint unsupervised learning of deep representations and image clusters," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5147–5156.

[41] W. Chu and D. Cai, "Stacked similarity-aware autoencoders," in *International Joint Conference on Artificial Intelligence*, 2017, pp. 1561–1567.

[42] X. Wang, K. Tieu, and E. Grimson, "Learning semantic scene models by trajectory analysis," in *European Conference on Computer Vision*, 2006, pp. 110–123.

[43] T. Zhang, H. Lu, and S. Z. Li, "Learning semantic scene models by object classification and trajectory clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1940–1947.

[44] X. Cui, Q. Liu, M. Gao, and D. N. Metaxas, "Abnormal detection using interaction energy potentials," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2011, pp. 3161–3167.

[45] F. Jiang, J. Yuan, S. A. Tsaftaris, and A. K. Katsaggelos, "Anomalous video event detection using spatiotemporal context," *Computer Vision and Image Understanding*, vol. 115, no. 3, pp. 323–333, 2011.

[46] M. Bertini, A. Del Bimbo, and L. Seidenari, "Multi-scale and real-time non-parametric approach for anomaly detection and localization," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 320–329, 2012.

[47] A. Adam, E. Rivlin, I. Shimshoni, and D. Reinitz, "Robust real-time unusual event detection using multiple fixed-location monitors," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 3, pp. 555–560, 2008.

[48] T. Xiang and S. Gong, "Video behavior profiling for anomaly detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 5, pp. 893–908, 2008.

[49] J. Kim and K. Grauman, "Observe locally, infer globally: a space-time mrf for detecting abnormal activities with incremental updates," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2921–2928.

[50] R. Mehran, A. Oyama, and M. Shah, "Abnormal crowd behavior detection using social force model," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 935–942.

[51] S. Wu, B. E. Moore, and M. Shah, "Chaotic invariants of lagrangian particle trajectories for anomaly detection in crowded scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2054–2060.

[52] B. Antić and B. Ommer, "Video parsing for abnormality detection," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2415–2422.

[53] V. Saligrama and Z. Chen, "Video anomaly detection based on local statistical aggregates," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2112–2119.

[54] H. Ren, W. Liu, S. I. Olsen, S. Escalera, and T. B. Moeslund, "Unsupervised behavior-specific dictionary learning for abnormal event detection." in *British Machine Vision Conference*, 2015, pp. 28–1.

[55] Y. Benezeth, P.-M. Jodoin, V. Saligrama, and C. Rosenberger, "Abnormal events detection based on spatio-temporal co-occurences," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 2458–2465.

[56] V. Mahadevan, W. Li, V. Bhalodia, and N. Vasconcelos, "Anomaly detection in crowded scenes," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 1975–1981.

[57] Y. Yang, Z. Ma, A. G. Hauptmann, and N. Sebe, "Feature selection for multimedia analysis by sharing information among multiple tasks," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 661–669, 2013.

[58] X. Zhu, Z. Huang, J. Cui, and H. T. Shen, "Video-to-shot tag propagation by graph sparse group lasso," *IEEE Transactions on Multimedia*, vol. 15, no. 3, pp. 633–646, 2013.

[59] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[60] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[61] Y. Yang, Z.-J. Zha, Y. Gao, X. Zhu, and T.-S. Chua, "Exploiting web images for semantic video indexing via robust sample-specific loss," *IEEE Transactions on Multimedia*, vol. 16, no. 6, pp. 1677–1689, 2014.

[62] J. K. Dutta and B. Banerjee, "Online detection of abnormal events using incremental coding length." in *AAAI Conference on Artificial Intelligence*, 2015, pp. 3755–3761.

[63] L. Zhu, Z. Xu, Y. Yang, and A. G. Hauptmann, "Uncovering the temporal context for video question answering," *International Journal of Computer Vision*, vol. 124, no. 3, pp. 409–421, 2017.

[64] L. Zhu, Z. Xu, and Y. Yang, "Bidirectional multirate reconstruction for temporal modeling in videos," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2653–2662.

[65] R. Tudor Ionescu, S. Smeureanu, B. Alexe, and M. Popescu, "Unmasking the abnormal events in video," in *IEEE International Conference on Computer Vision*, 2017, pp. 2895–2903.

[66] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. A. Efros, "Context encoders: Feature learning by inpainting," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2536–2544.

[67] D. Li, W.-C. Hung, J.-B. Huang, S. Wang, N. Ahuja, and M.-H. Yang, "Unsupervised visual representation learning by graph-based consistent constraints," in *European Conference on Computer Vision*. Springer, 2016, pp. 678–694.

[68] M. Aharon, M. Elad, and A. Bruckstein, "rmk-svd: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[69] X. Zhang, F. Zhou, Y. Lin, and S. Zhang, "Embedding label structures for fine-grained feature representation," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 1114–1123.

[70] P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, vol. 11, no. Dec, pp. 3371–3408, 2010.

[71] K. Soomro, A. R. Zamir, and M. Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.

[72] Y. S. Chong and Y. H. Tay, "Abnormal event detection in videos using spatiotemporal autoencoder," in *International Symposium on Neural Networks*. Springer, 2017, pp. 189–196.

[73] H. Vu, T. D. Nguyen, A. Travers, S. Venkatesh, and D. Phung, "Energy-based localized anomaly detection in video surveillance," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2017, pp. 641–653.

[74] M. J. Roshtkhari and M. D. Levine, "An on-line, real-time learning method for detecting anomalies in videos using spatio-temporal compositions," *Computer Vision and Image Understanding*, vol. 117, no. 10, pp. 1436–1452, 2013.

Wenqing Chu received his B.E. degree in Computer Science and Technology from Huazhong University Of Science And Technology, in 2014. He is currently a Ph.D. candidate in the State Key Laboratory of CAD&CG, College of Computer Science at Zhejiang University, China. His research interests include machine learning, computer vision and data mining.

Hongyang Xue received the B.E. degree in Computer Science and Technology from Zhejiang University, China, in 2014. He is currently a Ph.D. candidate in Computer Science at Zhejiang University. His research include machine learning, computer vision and data mining.

Chengwei Yao is a Lecturer in the College of Computer Science and Technology at Zhejiang University, China; He received the Ph.D. degree in Computer Science from Zhejiang University, China, in 2017. His research interests include machine learning, information retrieval and knowledge discovery.

Deng Cai is a Professor in the State Key Laboratory of CAD&CG, College of Computer Science at Zhejiang University, China. He received the Ph.D. degree in Computer Science from the University of Illinois at Urbana Champaign in 2009. His research interests include machine learning, data mining and information retrieval.