Crowd activity recognition in live video streaming via 3D-ResNet and region graph convolution network

Junpeng Kang | Jing Zhang | Wensheng Li | Li Zhuo

1 Faculty of Information Technology, Beijing University of Technology, Beijing, China
2 Beijing Key Laboratory of Computational Intelligence and Intelligent System, Beijing University of Technology, Beijing, China

Abstract
Since the era of we-media, live video industry has shown an explosive growth trend. For large-scale live video streaming, especially those containing crowd events that may cause great social impact, how to identify and supervise the crowd activity in live video streaming effectively is of great value to push the healthy development of live video industry. The existing crowd activity recognition mainly uses visual information, rarely fully exploiting and utilizing the correlation or external knowledge between crowd content. Therefore, a crowd activity recognition method in live video streaming is proposed by 3D-ResNet and regional graph convolution network (ReGCN). (1) After extracting deep spatiotemporal features from live video streaming with 3D-ResNet, the region proposals are generated by region proposal network. (2) A weakly supervised ReGCN is constructed by making region proposals as graph nodes and their correlations as edges. (3) Crowd activity in live video streaming is recognised by combining the output of ReGCN, the deep spatiotemporal features and the crowd motion intensity as external knowledge. Four experiments are conducted on the public collective activity extended dataset and a real-world dataset BJUT-CAD. The competitive results demonstrate that our method can effectively recognise crowd activity in live video streaming.

1 | INTRODUCTION

Since the era of we-media, live video streaming has become one of the most appealing Internet products [1]. With the typical characteristics of 'watching and playing as you go', the content of live video streaming becomes more and more diverse and complex. Especially, some live crowd events, with wide audience, fast spread, disordered content and difficult supervision, are a new battlefield of content governance in cyberspace. How to apply intelligent analysis techniques to identify and supervise the crowd activity in live video streaming is of great value to push the healthy development of live video industry to an even higher stage [2].

Similar to crowd counting, crowd tracking, density estimation or other crowd tasks, crowd activity recognition has a complex spatial–temporal relationship with the characteristics of the multiple groups that all have wide-spread applications including crowd monitoring, public safety and so on [3, 4]. Moreover, crowd activity recognition focuses on the crowd motion trend, the correlation of crowd and the description of the crowd activity [5, 6]. Since the large-scale crowd behaviour or movement has serious mutual occlusion, complex illumination changes and chaotic movement pattern, it makes the conventional analysis method for crowd tasks become unreliable, especially in high-density crowd state. Some effective approaches are still proposed to solve these challenges [7]. Early crowd activity recognition mainly relies on the combination of handcrafted features and classifiers. For example, Xu et al. [8] designed a method to detect abnormal crowd activity in complicated scenes by comprehensively considering both global and local spatiotemporal contexts. Due to shallow features leading to weak discrimination, the existing traditional approaches are trivial for the complex processing requirements of live video applications. Lately, thanks to the innovation of deep learning techniques, convolutional neural network (CNN) and recurrent neural network (RNN) as the representative deep networks were proposed for crowd activity detection and recognition and achieved good results. Avinash et al. [9] constructed a deep CNN model.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2021 The Authors. IET Image Processing published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology
by stochastic gradient descent whale optimization algorithm to detect the fighting, walking, escaping and other crowd activities, and obtained a high accuracy. Fan et al. [10] proposed an efficient spatiotemporal CNN to detect abnormal crowd activity in real time. Ravanbakhsh et al. [11] addressed the abnormality detection problem in crowded scenes by training a generative adversarial nets (GANs) with normal frames and corresponding optical-flow images to further improve classification speed and accuracy for behaviour videos. Xie et al. [12] replaced the 3D convolutions by low-cost 2D convolutions, obtaining an ideal result. Gupta et al. [13] extracted motion and appearance feature representations by deep convolution neural network (DCNN) to detect global abnormal motion in videos. Li et al. [14] proposed a deep end-to-end approach based on the long short-term memory (LSTM), showing a competitive performance in pedestrian path prediction and crowd behaviour classification. Due to the limitations of CNN’s convolution composition or RNN’s model structure, these deep networks can only deal with structured data (such as visual features of a picture), while ignore the unstructured data (such as correlation between features) [15], which is not suitable for complex crowd activities. Studies have shown that humans have remarkable ability to utilize acquired knowledge to reason about the dynamically changing world [16, 17]. When humans observe a crowd activity, they will not only pay attention to the people or objects in the scene, but also fully consider the correlation between them (as an external knowledge), to get a more accurate judgment. This fact shows the correlations between features are indispensable important clues for learning and reasoning [16]. Unfortunately, the above methods rarely consider the external knowledge that is one of the most important abilities for human beings to make an accurate judgment in certain scene [17].

In fact, the existing deep learning approaches can be further optimized by utilizing the correlation between visual features or constructing the network with multiple forms of data to apply in crowd activity recognition. More recent, graph neural networks (GNN), which can take advantage of unstructured data, has achieved great success in the field of computer vision, such as object detection, multi-label image classification, video reasoning. The current graph networks can be divided into supervised learning and non/weakly supervised learning. The graph networks with supervised learning may acquire nice results at the cost of huge resource consumption. Especially in the crowd activity recognition, there are some remarkable works, such as Yan et al. [4] designed a hierarchical graph-based cross inference network (HiGGCN) with three levels of information, that is, the body-region level, person level, and group activity level are learned and inferred in an end-to-end manner. Tang et al. [6] developed a semantics-preserving teacher–student (SPTS) network architecture with two graph convolutional modules, reaching a series of superior performances in the experiments. Wu et al. [18] built a flexible and efficient actor relation graph (ARG) to simultaneously capture the appearance and position relation between actors as well as the discriminative relation information for group activity recognition. While the results of graph networks using non/weakly supervised learning can also meet the general requirements with less resource consumption. These are convenient for algorithm deployment. Gao et al. [19] constructed an end-to-end video classification model based on a structured knowledge graph. The graph model can not only identify local knowledge structures in each video shot but also model dynamic patterns of knowledge evolution across these shots. Besides that, Gao et al. [20] further presented a novel graph convolutional tracking (GCT) by jointly incorporating two types of GCNs into a Siamese framework for target appearance modelling to greatly improve the stability of visual tracking. From the above works, we can see the graph network learning itself has many parameters and will consume huge computing cost, especially supervised learning that it is not flexible for many practical applications. Therefore, we intend to realize a weakly supervised learning for graph network to crowd activity.

For crowd activity in live video streaming, the visual focus is generally human centred, and the crowd is often moving. Considering that 3D-ResNet can obtain the visual features of each frame by 3D convolution operation as well as learn the changes between adjacent frames over time [21], we will use 3D-ResNet to extract the deep spatiotemporal features from live video. In the crowd activity of live video streaming, the whole scene of every frame is commonly divided into several regions, and the movement of each part of a crowd is tightly related. To make use of the visual features of each region in the video frames and the topological relationship between these features, we improve GCN to regional graph convolution network (ReGCN) to learn the correlation between data, that is, external knowledge, to promote the recognition performance. Besides that, crowd motion intensity (CMI) also as an external knowledge plays an important role in distinguishing crowd activities, and can participate in deep learning network to help boost recognition performance. Therefore, the CMI is utilized as one kind of external knowledge. Based on the above analysis, we propose a crowd activity recognition method in live video streaming via 3D-ResNet and ReGCN. First, the region proposals are generated by region proposal network (RPN) after extracting the deep spatiotemporal features from the live video streaming by 3D-ResNet. Then, a weakly supervised ReGCN is constructed by making region proposals as graph nodes and the correlation as edges. Finally, crowd activity in live video streaming is recognised by combining the output of ReGCN, the deep spatiotemporal features and the CMI as external knowledge.

The main contributions in this paper are introduced as follows:

1. Based on the region proposals in video frames generated by RPN, and the correlation between each region, we improve the original graph convolution network to construct a weakly supervised ReGCN. Our weak supervision is reflected in that the features of each region in the video frames are not finely classified, in which the training data are labelled roughly. These regions’ features are aggregated together to generate the video’s category.

2. The external knowledge in the crowd activity recognition is introduced to enrich the feature spaces. The external knowledge includes the deep spatiotemporal features extracted by 3D-ResNet and the intensity of crowd movement because
of its significant identification information for crowd activity recognition or crowd event detection.

The remainder of this paper is organised as follows. Section 2 mainly introduces the details of our method. Section 3 presents the experimental results and analysis. Conclusions and future work are discussed in Section 4.

2 METHODOLOGY

To effectively realise the crowd activity recognition in live video streaming, we construct a recognition network by combining 3D-ResNet and ReGCN, shown in Figure 1. First, the deep spatiotemporal features are extracted by 3D-ResNet, and then the region proposals in the video frames are generated by RPN [22]. After obtaining the region proposals of the video, a weakly supervised ReGCN with 2–6 layers is constructed by making these region proposals as graph nodes and their correlations as edges. Considering the nodes’ features of ReGCN are only the simple combination information of each region proposals generated by RPN, and the intensity of crowd motion are also different in various types of crowd activities, we obtain the intersection over union (IoU) information and relative position information by comparing these region proposals. Finally, crowd activity in live video streaming is recognised by combining the deep spatiotemporal features extracted by 3D-ResNet and the CMI as external knowledge. Here, the CMI and deep spatiotemporal features are taken as the external knowledge to improve the discriminative ability of our model.

2.1 Deep spatiotemporal feature extraction and region proposal generation

ResNet is one of the most powerful architecture and considering 3D Convnets are more suitable for spatiotemporal features [23]. Since the architecture of 3D-ResNet is expected to achieve better performance than relative models, we use 3D-ResNet to extract the deep spatiotemporal features from live video streaming.

First, we need to resize live video streaming to capture the key frames. In this paper, we select a live video streaming for 5–10 s, capture the key frame at 6 frames per second (FPS), and take 32 consecutive key frames to form clip to represent one live video streaming. These key frames are resized to 224 × 224 and sent into 3D-ResNet to extract the deep spatiotemporal features, in which the depth of 3D-ResNet is 50, and the residual modules is 16 in total. The model structure is shown in Table 1 [23], in which the size of convolution kernels is 5 × 7 × 7 for the Conv1 and the other convolution kernels are given in the middle column, the temporal stride of Conv1 is 1, the size of input clips is 3 × 224 × 224, and the output size of Conv1 is 32 × 112 × 112. Specially, downsampling of the inputs is performed by Conv2, Conv3, Conv4, Conv5 with a stride of 2.

By 3D-ResNet, video frames are finally transformed into a feature matrix with the size of $T \times H \times W \times d$ (where $T$ is the video frame sequence, $H \times W$ is the size of the feature map and $d$ represents the number of channels). This feature matrix
is used for subsequent crowd activity recognition through two branches. One branch is the average pooling of deep spatiotemporal features extracted by 3D-ResNet, which could generate one $d$-dimensional vector. In the fully connected layer, the $d$-dimensional vector is input as an external knowledge of live video streaming for recognising crowd activity.

The other branch is to generate region proposals based on RPN derived from fast region-based convolutional network (Fast R-CNN) by utilizing the feature maps from the convolution layer of 3D-ResNet to generate the initial region proposals. The region proposed by RCNN is pretrained on the ImageNet to fine tune the bounding box to detect certain objects. Since the crowd activity in live video streaming may cause errors when accurately locating certain objects with bounding box regression, and the computation cost of the prediction box from the bounding box is very heavy, we remove the bounding box regression from the original RPN to roughly get the information of some regions. Then, hyperparameter $N$ region proposals are generated through the sliding window. As shown in Figure 2, in every channel of $T$ frame convolution feature map, furthermore, $d$ channels can generate $N \times d$ region proposals as graph nodes in following ReGCN.

### 2.2 | ReGCN construction by region proposals and their correlations

In this section, we will construct a weakly supervised ReGCN by using the region proposals from every channel. Our ReGCN is constructed by making region proposals as graph nodes and their correlations as edges. These region proposals’ features in video frames are not fine classified. The exact labels of training data are given only on the whole clips of videos and the features of region proposals are roughly aggregated together in our weakly supervised methods. The correlations in our graph network are computed by similarity between region proposals. Therefore, the graph representation and graph learning of crowd activity are realised by ReGCN. More concretely, the spatial features of our region proposals topological graph can be extracted by ReGCN, optimizing and updating the state information of every node in the graph structure is also executed to realise the convolution calculation of non-Euclidean data [24]. Compared with the traditional deep learning method like CNN, our model includes not only the inherent attributes of feature nodes, but also the graph structure information for describing the relationship between different features. ReGCN can utilize the attribute information of nodes and the graph structure information between nodes in the same graph convolution layer, so that the result of our model can be cooperatively influenced by the different features [25]. As shown in Figure 3, the correlation between different region proposals in crowd activity is described as a kind of similarity relationship between two region proposals, which can generate high confidence edges in the graph structure. Based on [26] and [27], the correlations graph convolution structure with 2–6 layers is constructed.

The basic form of graph convolution module is denoted as:

$$
g_{\theta} \odot x \approx \theta \left( I_M + D^{-\frac{1}{2}} A \cdot D^{-\frac{1}{2}} \right) x, \quad (1)$$

where $g_{\theta}$ is the diagonal matrix composed of Fourier transform, $\odot$ is a multiplication operation between matrices of the same order, $x$ is the input of graph convolution, $\theta$ is the parameter in the convolution kernel of Chebyshev polynomial, $I_M$ is the information of each node, $M$ represents all the nodes in the graph structure, $D$ is the degree matrix of nodes, and $A$ is the adjacency matrix of nodes. Considering that the repeated stacking of (1) in a deep graph network may lead to the disappearance or explosion of the gradient, (1) is normalised to:

$$
Z \approx \tilde{D}^{-\frac{1}{2}} A \cdot \tilde{D}^{-\frac{1}{2}} X \Theta, \quad (2)
$$

where $\tilde{A} = A + I_N$ means the adjacency matrix with self-loop connection of undirected graph, $\tilde{D} = \sum_j \tilde{A}_{ij}, X$ is the aggregate of input signal $x$ in different channels, $\Theta$ represents the param-

---

**FIGURE 2** The region proposal generation in live video

**FIGURE 3** Construction of our ReGCN
eter weight matrix of convolution kernel, and $Z$ represents the output matrix of normalised convolution equal to the output characteristics of each convolution layer.

The nodes in our ReGCN are the region proposals extracted from video frames by 3D-ResNet and RPN. The edges of the graph structure are the similarity relationships between the region proposals. The graph structure can be expressed as $G = (V, E, W)$, in which, $V = \{x_1, x_2, ..., x_N\}$ represents the $N$ region proposals in the video frame, $E$ represents the correlations between the region proposals, and $W$ represents the learnable weight matrix. The feature similarity $S$ between two region proposals $i$ and $j$ is described as:

$$S(x_i, x_j) = b(x_i)^T b'(x_j),$$

where $b(x_i) = W \times x_i$, $b'(x_j) = W' \times x_j$, $W$ and $W'$ are the learnable weight matrix of $d \times d$ dimension, respectively. The similarity between regions of video frame can be calculated by (3). The value of each edge is standardised so that the sum is 1. The normalised values can form the adjacency matrix of the graph.

By substituting the adjacency matrix composed by (3) into (2) of graph convolution network, we can get that:

$$Z = S \cdot X \cdot W,$$

where $Z$ is the output feature graph of convolution layer, the size of which is $N \times d$, $S$ is the adjacency matrix of graph nodes, the size of which is $N \times N$; $X$ is the $N \times d$ vector composed of $N$ regional features extracted from each video frame, which is the node input information of graph convolution network; $W$ is the size $d \times d$ weight matrix that can be learned by back propagation of graph convolution network. Based on other forms of adjacency matrix, such as the adjacency matrix with distance, they may result in many 0 values, which will interrupt the program in the actual implementation process. Therefore, we adopt the adjacency matrix based on similarity. So far, the correlations graph convolution structure has been constructed.

2.3 Crowd activity recognition by combining ReGCN with deep spatiotemporal features and crowd motion intensity

In the previous section, ReGCN was constructed by the nodes’ features of region proposals or the correlations between them. As a result, the input of ReGCN only covers the features of partial region proposals. The single source information is not conducive to the network recognition performance. Considering the intensity of crowd movement is also an important identification information for crowd activity recognition or crowd event detection, we will introduce external knowledge including the CMI and deep spatiotemporal features to raise recognition performance. Thus, our ReGCN model takes the graph convolution network as the benchmark and then external knowledge is connected to participate in crowd activity recognition.

Since the deep spatiotemporal features extracted by 3D-ResNet are discussed in Section 2.1, we mainly focus on CMI.

In the analysis of crowd movement, we adopt the movement intensity to represent the movement of video clips. Considering that the computation cost of inter frame difference information or optical flow, and these features are easy to be affected by noise, we use the region proposals generated by RPN to roughly describe the average intensity of crowd movement. The size of the region proposals and its displacement distance between different frames are calculated.

As shown in Figure 4, we describe the matching relationship by the overlap degree of region proposals from adjacent frames. To prevent highly relevant content that leads to the mis-looking at the irrelevant background, and also to avoid weakly relevance content that makes proposals hard to match each other, the IoU feature is taken as 0.5–0.7 between the region of frame $t$ and the region of frame $t + 1$. Using the area of the motion of region proposals and the displacement distance of the centre point, the average intensity of motion is calculated.

$$B = \frac{1}{T - 1} \sum_{t=1}^{T-1} \left( \frac{1}{N} \sum_{i=1}^{N} K_i L_i^2 \right),$$

where $B$ is the average crowd intensity of video clips, $T$ is the number of video frames, $N$ is the number of region proposals in frame $t$, $K$ is the area of every region, and $L$ is the displacement distance of the centre point of the regions between two frames. The average motion intensity between frames $t - 1$ and $t$ can be seen in brackets. For the whole video clip, the average motion intensity $B$ is the average of $(T - 1)$ video frames.

The CMI $B$ is normalised and expanded to a $(1 \times d)$-dimensional vector. The output features of the last graph convolution layer and the deep spatiotemporal features are both averaged pooled. Thus, crowd activity in live video streaming is recognised by combining the three branches (ReGCN, deep spatiotemporal features and CMI) together into the final full connection layer.

3 EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental setup

We evaluate the performance on a public dataset collective activity extended data (CAE) [28] and a self-collected dataset.
BJUT-CAD (crowd activity dataset). The CAE dataset is the extension of collective activity data (CAD) dataset [29]. In total, CAE dataset consists of 6 classes and 23,431 frames, which is divided into crossing, waiting, talking queueing, dancing and jogging. Several samples in CAE are shown in Figure 5.

The other dataset is our BJUT-CAD collected from various live streaming platforms and video websites. BJUT-CAD has 830 short videos, in which the main content is crowd activity including eight types of square dance, bands march, sports meeting, tourist street roaming, military parade, small-scale fighting, demonstration and large-scale riot. On average, there are 100 short videos ranging 5–10 s in each category. The key frames are taken in the speed of 6 FPS, and every video takes its continuous 32 frames. The several samples of BJUT-CAD are shown in Figure 6.

Experiments are conducted with an Intel Xeon E5-2620, 16 GB RAM, and an Ubuntu 20.04 operating system. Our framework is implemented by PyTorch, accelerated by a GeForce RTX 2080Ti device with 11 GB memory, CUDA 10.2 and cuDNN 440.95.01. The proposed method is implemented on CAE and BJUT-CAD dataset, respectively. We acquired the training set and test set with a rate of 7:3. During training process, each frame is resized to 224 × 224 and then fed into 3D-ResNet50 to obtain regional features. This model is trained using SGD with a learning rate of 7.5 × 10⁻⁵ and 100 epochs. Moreover, we set the impulse to 0.9, and dropout to 0.3. After the 85th epoch, the learning rate will be adjusted to 7.5 × 10⁻⁶ to avoid overfitting. After obtaining the information of regional feature nodes, these feature nodes are sent into ReGCN with 1.25 × 10⁻⁴ initial learning rate. The learning rate will be also adjusted to 0.1 times of the original value after the 85th iteration.

In our experiments, recognition accuracy is adopted as the performance criterion, which is computed by evaluating the correct predictions out of the total predictions as:

\[
 r = \frac{TP + TN}{TP + TN + FP + FN}\tag{6}
\]
where $TP$ represents the true positives, $TN$ represents the true negatives, $FP$ denotes the false positives, and $FN$ stands for the false negatives. In addition, the FPS is used to measure the recognition speed, which represents the number of images that the recognition model can process per second.

### 3.2 Experiment I: The influence of different ReGCN parameters on accuracy

Because different graph convolution layers and nodes number might influence the recognition performance, we conduct two groups of experiments with various relevant parameters in BJUT-CAD dataset. The experimental results are shown in Figures 7 and 8.

As can be seen in Figure 7, we compare different ReGCN layers on different datasets. The ReGCN with two graph convolution layers achieves the highest accuracy of 90.20% on CAE dataset and 86.75% on BJUT-CAD dataset, respectively. While the accuracy of 6 layers is the lowest in several layers, which is 82.35% and 82.53% on the two datasets, respectively. Except layer 1 of ReGCN, the accuracy is gradually decreasing with the increase of the ReGCN layers. In general, more detailed features can be obtained that are helpful for crowd activity recognition when increasing the number of layers. But in view of the over smoothing phenomenon in graph convolution, the result may not be a linear change. From the perspective of the frequency domain, graph convolution in ReGCN is equivalent to a low-pass filter, which can learn and filter the input graph nodes signals. With the operation of stacking, the frequency response of graph convolution will be more sensitive to the low-frequency input of the graph nodes signals. Generally, the low-frequency signals are more related to task targets, while the high-frequency signals contain more noise. This low-pass filtering property of graph convolution in ReGCN will make the smoothing phenomenon gradually decrease with the increase of ReGCN layers.

From the experimental results, we can see that a deeper graph convolution structure does not necessarily contribute to recognition performance, and the two-layer ReGCN is more optimal.

To explore the effect of nodes numbers, we compare the recognition accuracy with different number of nodes, shown in Figure 8, in which ReGCN starts with $5d$ graph nodes and gradually increases to $30d$ nodes ($d$ is the number of channels).

We can see that the accuracy of ReGCN achieves the highest recognition accuracy of 90.20% on CAE dataset when the number of nodes is $10d$, and 86.75% is acquired on BJUT-CAD when the number of nodes is $15d$. The performance of the network model will decline whether the nodes’ numbers increase or decrease. We consider that graph structure itself is a high-dimensional sparse data structure. For the graph network with $N \times d$ feature nodes ($N$ is the number of region proposals), if there are too few graph nodes, the ReGCN could not obtain enough node features and the correlations between nodes. When $N$ is too large, that means we extract too many region proposals in the same frame. This will cause much repeated extraction for ReGCN inputs. If the graph data in ReGCN cannot meet the independent distributed, many repetitive nodes will bring redundancy and noise, which is not conducive to the subsequent processing. We also discuss the difference between the best amount of these two datasets. For CAE dataset, we apply $10d$ graph nodes while applying $15d$ on BJUT-CAD. This difference may be caused by the complexity of the content in video frames. As we can see from the several samples of each dataset, the content of BJUT-CAD is more complex and disordered than CAE. Therefore, we need to extract more graph nodes on BJUT-CAD.

### 3.3 Experiment II: Ablation study of different modules

To validate the effect of different modules in proposed method, we conduct a group of ablation study in CAE and BJUT-CAD datasets. Based on the results of Experiment I, the pre-processing network of ReGCN is 3D-ResNet50, and nodes are set to 10 and 15 on the CAE and BJUT-CAD datasets, respectively. The experimental results are shown in Figure 9, including three modules: deep spatiotemporal feature extraction (DSTF), ReGCN and CMI module. From Figure 9, we can see the recognition accuracy of ReGCN is 64.71% on CAE and 62.75% on BJUT-CAD. The results show that if the graph convolution network is used alone, the recognition rate is underperformance. Because the graph node features and correlation features do not include the position information of each node in the video frames. And the scenes of the crowd and other information are also lost. Then, ReGCN + DSTF obtains the
accuracy of 86.27% and 81.23% on each dataset. The DSTF model is composed of the convolutional features of the last convolution layer in 3D-ResNet including the spatiotemporal features and high-level semantic features. That can greatly promote the network to improve performance. In addition, we also test the effect of ReGCN + CMI. In this experiment, the accuracy is 66.67% on CAE dataset and 68.67% on BJUT-CAD. Compared with using ReGCN alone, the performance improvement is not obvious. It may be that CMI is mainly related to the IoU information between different proposal regions while it is not learnable, therefore the improvement is limited. Only when integrating the three modules, it can capture multiple features, achieving the best performance of 90.20% and 86.75% on each dataset, respectively. The difference of the total accuracy on BJUT-CAD and CAE is because that the CAE dataset is constructed in a relatively stable environment by handheld photographic equipment. While the BJUT-CAD dataset is collected from various live streaming platforms and video websites, in which the videos are uploaded by people from all over the world. The BJUT-CAD dataset has more different details like various behaviours and the background scenes are more complex than CAE dataset. Even some frames or types of videos in the two datasets are totally different. We can find several samples in Figures 5 and 6. The people in CAE dataset are dancing or running in the cleaning and simple scenes in Figure 5 while the baseball player in BJUT-CAD dataset is catching the ball in a complex scene with many audiences in Figure 6. Especially, there are six kinds of videos in CAE dataset, which are divided into crossing, waiting, talking, queuing, dancing and jogging. While BJUT-CAD includes eight types of crowd activities such as square dance, bands march, sports meeting, tourist street roaming, military parade, small-scale fighting, demonstration and large-scale riot. We think that more kinds of videos will make model confusing and learning less powerful. Therefore, the accuracy on BJUT-CAD dataset is lower than the CAE.

3.4 Experiment III: Subjective results of crowd activity

To intuitively demonstrate the effectiveness of the region selection (graph nodes), we give the subjective results of ReGCN on BJUT-CAD datasets. In this paper, the attention heat map is generated on part of the experimental results by using the output of the convolution layer in 3D-ResNet and the visualization mechanism of Grad-CAM++ [30], shown in Figure 10, in which the red regions represent the visual focus. The visualization result Figure 10a2 is the heat map of Figure 10a1. Our model focuses on the centre of a group of people who are dancing. This is because the overall action of the crowd is the key factor for the model to judge the crowd activity category. Comparing these two pictures, the selection of region is more consistent with the main concerns of the model. From Figure 10b2, the model also focuses on the objects in the lower right corner. This shows that exercise intensity and group action characteristics are more important to judge the model. We can see a man is hitting the others in Figure 10c1 while the attention in Figure 10c2 is concentrating on this corner of chaos. As shown in Figure 10d2, obviously, the model captures the musicians holding trombones of Figure 10d1, in which the musicians holding almost invisible small musical instruments at the centre of the lens are ignored by the model. Maybe the model that musical instruments and other items are the key to identify whether the current group is a band. In addition, this paper also gives some examples of mismatches. A baseball pitcher is throwing out a baseball in Figure 10e1, and because of the obvious court boundary in Figure 10e2, the attention region is mainly on the pitcher with square edges. In Figure 10f2, the model not only finds the dancing action of the human body in Figure 10f1, but also the black border of the picture on the flagpole and the logo of the dance team in the lower right corner. This may because some dance video samples in the training data of this model have black borders or the same dance team logo, which makes the focus of the model not match the expectation.

3.5 | Experiment IV: Performance comparison with state-of-the-art methods

In this experiment, we comprehensively compare the recognition accuracy and speed with the state-of-the-art methods on the CAE dataset and BJUT-CAD dataset. The comparison methods include 3D-ResNet34 [31], ResNet(2 + 1)D [32], 3D-ResNet50 [31], WideResNet [33], LRCN [34] and VGG16 [35]. Among them, we use 3D-ResNet50 and WideResNet as backbone, respectively, shown in Table 2.

From the Table 2, we can see that ours (3D-ResNet50) obtains 86.75% accuracy on BJUT-CAD dataset, 90.20% on CAE dataset. 3D-ResNet34 obtains 74.10% accuracy on BJUT-CAD dataset. ResNet (2 + 1) D obtains the accuracies of 79.52% accuracy on BJUT-CAD dataset, 87.34% on CAE dataset. 3D-ResNet50 obtains 80.72% accuracy on BJUT-CAD dataset, 84.31% on CAE dataset. WideResNet accuracy achieves 83.73% on BJUT-CAD dataset. Ours (WideResNet) obtains 85.54% on BJUT-CAD dataset. LRCN obtains the accuracies of 69.79% on CAE dataset. VGG16 achieves 77.63%
FIGURE 10  Some examples of attention heat map, in which a1, b1, c1, d1, e1, f1 are the original video frames, and a2, b2, c2, d2, e2, f2 are the corresponding attention heat maps.

FIGURE 11  Performance comparison results of different methods

TABLE 2  Performance comparison with state-of-the-art methods

| Dataset | Network         | FPS  | Overall accuracy (%) |
|---------|-----------------|------|----------------------|
| BJUT-CAD| 3D-ResNet34 [31]| 131.69 | 74.10                |
|         | ResNet (2+1)D [32]| 122.14 | 79.52                |
|         | 3D-ResNet50 [31]| 125.99 | 80.72                |
|         | WideResNet [33]  | 120.76 | 83.73                |
|         | Ours (WideResNet)| 14.08  | 85.54                |
|         | Ours (3D-ResNet50)| 31.37  | 86.75                |
| CAE     | LRCN [34]       | —     | 69.79                |
|         | VGG16 [35]      | —     | 77.63                |
|         | 3D-ResNet50     | 168.87 | 84.31                |
|         | ResNet (2 + 1)D | 172.74 | 87.34                |
|         | Ours (3D-ResNet50)| 27.87  | 90.20                |

accuracy. At the same time, we can see ours (3D-ResNet50) can achieve better results on BJUT-CAD dataset which is superior to the performances of the 3D-ResNet34, ResNet(2 + 1)D, 3D-ResNet50, WideResNet and our method(WideResNet) by 12.61%, 7.23%, 6.03%, 3.02%, 1.21%, respectively. The accuracy of ours (3D-ResNet50) on CAE dataset is superior to the performances of LRCN, VGG16, 3D-ResNet50 and ResNet (2 + 1)D by 20.41%, 12.57%, 5.89%, 2.86%, respectively. In terms of recognition speed, although 3D-ResNet34 achieves 131.69 FPS on BJUT-CAD dataset, its accuracy is lower because the feature discriminative ability is not sufficiently robust. 3D-ResNet50 reaches 125.99 FPS, is a little slower than 3D-ResNet34. 3D-ResNet50 is the classic ResNet with 3D convolution kernels like 3D-ResNet34, but has deeper structure of 50. That is the more convolution layers produce the slower speed. WideResNet is 120.76 FPS slower than other ResNet, because of the wider network structure and the more parameters. As a complex structure of graph networks, ours (WideResNet) obtains 14.08 FPS on BJUT-CAD dataset and ours (3D-ResNet50) has different 31.37 and 27.87 FPS on two datasets, respectively. The graph convolution structure brings less FPS than traditional CNN methods, but ours (3D-ResNet50) obtains higher accuracy for the two datasets.

To present the performance of every model more intuitively, we take the FPS as the horizontal axis and the accuracy as the vertical axis and draw the model performance comparison as shown in Figure 11. We can clearly see that the nodes of 3D-ResNet50 and ResNet (2 + 1)D are in the right parts with 168.87
and 172.74 FPS while our methods are in the left positions with the 14.08, 27.87 and 31.37 FPS, respectively. However, the nodes of our method are higher than others with the accuracies of 86.75% on BJUT-CAD dataset and 90.20% on CAE dataset. That is, although some mainstream model methods are faster in speed, our ReGCN method has achieved higher accuracy in both BJUT-CAD dataset and CAE dataset. For the live video streaming, the 24 FPS can also meet the application demands. It shows our method can improve the recognition performance of crowd activity under the acceptable rate.

4 CONCLUSION AND FUTURE WORK

With the rapidly development of live video streaming, how to effectively identify and supervise the crowd activity from live videos has become a challenging task. This scenario aims to effectively recognise kinds of crowd activity in live video streaming. To realise our goals, a crowd activity recognition method in live video streaming is proposed via 3D-ResNet and ReGCN. First, we extract the deep spatiotemporal features of live video streaming by 3D-ResNet to generate region proposals by RPN. Then, we proposed to construct a weakly supervised ReGCN by region proposals as the graph nodes and the correlations between region proposals are used as edges. At last, the crowd activity in live video streaming is recognised by combining the output of ReGCN, the deep spatiotemporal features and the CMI as external knowledge to enrich the semantic information in the process of learning. The area and displacement information of each region between different frames are obtained by the IoU values and related calculation. Because there are no learnable parameters in this kind of external knowledge, the overall performance improvement of the network is weaker. Fortunately, the deep spatiotemporal characteristics extracted by 3D-ResNet are remarkable for crowd activity recognition. Expected the above work, we built a real-world video dataset called BJUT-CAD that includes eight kinds of crowd activity video collected from live video websites. To verify the effectiveness of our work, four experiments are conducted on BJUT-CAD and a public CAE dataset. In Experiment I, two kinds of optimal parameters in ReGCN are obtained by experimental demonstrations, respectively. At Experiment II, ablation study is conducted with different modules of our neural network. Due to the lack of various information features, it is not optimal to use graph convolution network alone. The recognition performance of the model will be improved significantly on combining the graph convolution network with the CMI and deep spatiotemporal feature of 3D-ResNet. In addition, we also compare the subjective results with several selection graph nodes in Experiment III, so that we can perceive the rationality of the selected graph nodes. Finally, Experiment IV compares with state-of-the-art crowd activity recognition methods on BJUT-CAD and CAE datasets, and a competitive result is obtained, proving the effectiveness of our method.

In the future, some lightweight work can be considered to GCN. We will modify our GCN to be lightweight to save computing resources, for example, teacher–student network. In addition, we need to further explore the non/weakly supervised graph network or model tailoring and network lightweight, to enhance the ability of graph network at the lower cost.

ACKNOWLEDGEMENTS

This work was supported in part by the National Natural Science Foundation of China under Grant 61971016, and in part by the Beijing Municipal Education Commission Cooperation Beijing Natural Science Foundation under Grant KZ201910005007.

ORCID

jingZhang https://orcid.org/0000-0003-1290-0738

REFERENCES

1. Camacho, D., Ma, V.L., Cambria, E.: New research methods & algorithms in social network analysis. Future Gener. Comput. Syst. 114, 290–293 (2021)
2. Wang, L., et al.: Porn streamer recognition in live video streaming via attention-gated multimodal deep features. IEEE Trans. Circuits Syst. Video Technol. 30(12), 4876–4886 (2020).
3. Wang, Q., et al.: Detecting coherent groups in crowd scenes by multiview clustering. IEEE Trans. Pattern Anal. Mach. Intell. 42(1), 46–58 (2020)
4. Yan, R., et al.: HiGCIN: hierarchical graph-based cross inference network for group activity recognition. IEEE Trans. Pattern Anal. Mach. Intell. 1–1 (2020), https://doi.org/10.1109/tpami.2020.3034233
5. Wang, Q., et al.: Pixel-wise crowd understanding via synthetic data. Int. J. Comput. Vision 129(1), 225–245 (2021)
6. Tang, Y., et al.: Learning semantics-preserving attention and contextual interaction for group activity recognition. IEEE Trans. Image Process. 28(10), 4997–5012 (2019)
7. Wang, Q., et al.: NWPU-crowd: a large-scale benchmark for crowd counting and localization. IEEE Trans. Pattern Anal. Mach. Intell. 1–1 (2020), https://doi.org/10.1109/tpami.2020.3013269
8. Xu, D., et al.: Video anomaly detection based on a hierarchical activity discovery within spatio-temporal contexts. Neurocomputing 143, 144–152 (2014)
9. Avinash, R.: Stochastic gradient descent whale optimization algorithm based deep convolutional neural network to crowd emotion understanding. Comput. J. 2(2), 267–282 (2020)
10. Fan, Z., et al.: Real-time and accurate abnormal behaviour detection in videos. Mach. Vis. Appl. 31(10), 7–8 (2020)
11. Ravanbakhsh, M., Nabi, M., Sangineto, E., Marcenaro, L., Regazzoni, C., Sebe, N.: Abnormal event detection in videos using generative adversarial nets. In: IEEE International Conference on Image Processing (ICIP), Beijing, pp. 1577–1581 (Sep. 2017).
12. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. In: Proceedings of the European Conference on Computer Vision (ECCV), Munich, pp. 305–321 (2018)
13. Gupta, T., Namavath, V., Roy, S.: Crowd VAS-Net: a deep-CNN based framework to detect abnormal crowd-movement behaviour in videos for predicting crowd disaster. In: IEEE International Conference on Systems, Man and Cybernetics (SMC), Bari, Italy, pp. 2877–2882 (2019)
14. Li, Y.: A deep spatiotemporal perspective for understanding crowd behaviour. IEEE Trans. Multimedia 20(12), 3289–3297 (2018)
15. Mabrouk, A., Zagrouba, E.: Abnormal behaviour recognition for intelligent video surveillance systems: a review. Expert Syst. Appl. 91(1), 480–491 (2018)
16. Gao, J., Zhang, T., Xu, C.: I know the relationships: zero-shot action recognition via two-stream graph convolutional networks and knowledge graphs. In: The AAAI Conference on Artificial Intelligence, Hawaii, pp. 8303–8311 (2019)
17. Sabokrou, M., et al.: Deep-anomaly: fully convolutional neural network for fast anomaly detection in crowded scenes. Comput. Vision Image Understanding 172(6), 88–97 (2018)
18. Wu, L., et al.: Learning actor relation graphs for group activity recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, pp. 9956–9966 (2019)
19. Gao, J., Zhang, T., Xu, C.: Watch, think and attend: end-to-end video classification via dynamic knowledge evolution modeling. In: ACM Multimedia Conference, Seoul, pp. 690–699 (2018)
20. Gao, J., Zhang, T., Xu, C.: Graph convolutional tracking. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, pp. 4644–4654 (2019)
21. Tran, D., et al.: Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, pp. 4489–4497 (2015)
22. Ren, S., et al.: Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans. Pattern Anal. Mach. Intell. 39(6), 1137–1149 (2017)
23. Hara, K., Kataoka, H., Satoh, Y.: Learning spatio-temporal features with 3D residual networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, pp. 3154–3160 (2017)
24. Schlichtkrull, M., et al.: Modeling relational data with graph convolutional networks. In: European Semantic Web Conference, pp. 593–607 (2018)
25. Gama, F., et al.: Graphs, convolutions, and neural networks: From graph filters to graph neural networks. IEEE Signal Process Mag. 37(6), 128–138 (2020)
26. Kipf, TN., Welling, M.: Semi-supervised classification with graph convolutional networks. : International Conference on Learning Representations (ICLR), Toulon, pp. 1–14 (2017)
27. Wang, X., Gupta, A.: Videos as space time region graphs. In: Proceedings of the European conference on computer vision (ECCV), Munich, pp. 399–417 (2018)
28. Choi, W., Shahid, K., Savarese, S.: Learning context for collective activity recognition. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Colorado, pp. 3273–3280 (2011)
29. Choi, W., Shahid, K., Savarese, S.: What are they doing? Collective activity classification using spatio-temporal relationship among people. In: IEEE International Conference on Computer Vision Workshops (ICCV), Kyoto, pp. 1282–1289 (2009)
30. Aditya, C., et al.: Grad-CAM+: generalized gradient-based visual explanations for deep convolutional networks. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, pp. 839–847 (2018)
31. He, K., et al.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, pp. 770–778 (2016)
32. Tran, D., et al.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, pp. 6450–6459 (2018)
33. Sergey, Z., Nikos, K.: Wide residual networks. In: Proceedings of the British Machine Vision Conference (BMVC), York, pp. 1–12 (2016)
34. Donahue, J., et al.: Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans. Pattern Anal. Mach. Intell. 39(4), 677 (2017)
35. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (ICLR), San Diego, pp. 1–14 (2015)

How to cite this article: Kang J, et al.: Crowd activity recognition in live video streaming via 3D-resnet and region graph convolution network. IET Image Process. 1–11 (2021). https://doi.org/10.1049/ipr.12239