Detecting Global Exam Events in Invigilation Videos Using 3D Convolutional Neural Network

Zichun Dai, Chao Sun(✉), Xinguo Yu, and Ying Xiang

National Engineering Research Center for E-Learning, Central China Normal University,
Wuhan, China
daizichun@mails.ccnu.edu.cn, {csun,xgyu}@mail.ccnu.edu.cn,
yingxiang333@outlook.com

Abstract. This paper designs a structure of 3D convolutional neural network to detect the global exam events in invigilation videos. Exam events in invigilation videos are defined according to the human activity performed at a certain phase in the entire exam process. Unlike general event detection which involves different scenes, global event detection focuses on differentiating different collective activities in the exam room ambiance. The challenges lie in the great intra-class variations within the same type of events due to various camera angles and different exam room ambiances, as well as inter-class similarities which are challengeable. This paper adopts the 3D convolutional neural network based on its ability in extracting spatio-temporal features and its effectiveness in detecting video events. Experiment results show the designed 3D convolutional neural network achieves an accuracy of its capability of 93.94% in detecting the global exam events, which demonstrates the effectiveness of our model.

Keywords: Exam event detection · Surveillance video · 3D convolutional neural network

1 Introduction

In the modern society the exams are the important activity because they are widely used to evaluate the individual ability. However, the traditional invigilation needs a large number of human resources, which is expensive. Another demerit of invigilating by human being is hardly to get rid of subjective judgment on exam events. The demerits of the traditional invigilation motivate us to develop automatic invigilation systems. An automatic invigilation model is proposed for detecting suspicious activities in exams [1]. Cote *et al.* applies two-state Hidden Markov model to distinguish the abnormal exam events from the normal exam events [2]. However, they take slight attention on global exam event detection. Global event detection is an essential and core part of a complete automatic invigilation system, which can be the prior task of abnormal exam event detection. Besides, the automatic detection of the global exam events benefits the communication between invigilation system and management system, which is convenient for further conducting examination evaluation and analysis. Furthermore, the fairness of exam can be improved by reducing the human subjective judgment.

J. J. Dabrowski et al. (Eds.): PSIVT 2019 Workshops, LNCS 11994, pp. 172–182, 2020.
https://doi.org/10.1007/978-3-030-39770-8_14

One goal of automatic invigilation system is to detect the exam events from the invigilation videos. However, video events detection is still a challenging task due to the background clutter or occlusions. The successful methods in the recent years focuses on extracting spatial-temporal features like STIP [3], HOG3D [4], MBH [5] and dense trajectories [6], and uses bag-of-visual-word histograms or Fisher vectors [7] to represent their distribution in a video for classification. Among these features, improved Dense Trajectories (iDT) perform the best. However, extracting these local features is time-consuming and some discriminative features make difference in finer part of the whole video.

With the breakthrough of image classification brought by convolutional neural network [8], recent researches concentrate on applying convolutional neural network to video events detection. Karpathy *et al.* firstly apply convolutional neural network with different time fusion strategies on Sports-1M video dataset but gain less accuracy than hand-crafted features [9]. Feichtenhofer *et al.* explore the two-stream convolutional network fusion for action recognition in videos. Interestingly, they find that the slow fusion of temporal and spatial network can boost accuracy of classification [10]. Although it is time-consuming by training two networks, it inspired researchers that the temporal information is critical for understanding the video activities. 3D convolutional neural network is then be proposed [11]. It builds an architecture to directly learn the spatial and temporal features by adding temporal dimension to the network. It performs well in extracting compact and discriminative features, which is necessary for efficient video event detection.

Global exam event detection in invigilation videos is a branch of video event detection. Different form detecting abnormal exam behavior like cheating, it aims at detecting the whole status in the exam room. For global exam events detection, one challenge lies in the great intra-class variations. The events take place in different classrooms and the classrooms are invigilated from various angles of cameras, which means the same exam event can occur with different background. Additionally, global exam events detection suffers from inter-class indistinguishability as a result of the finer motion change in different events [12]. It is hard to distinguish the event accurately from one still image. To the best of our knowledge, there is no research targeting at global exam event classification before. After reviewing the technologies in video event detection, we notice the efficiency of 3D convolutional neural network of extracting compact and discriminative features. Therefore, this paper develops the detecting model based on the 3D convolutional neural network to solve the detection problem for global exam events.

2 Proposed 3D Convolutional Neural Network Structure

We detect the global exam events using the 3D convolutional neural network. The proposed method firstly breaks each predefined exam surveillance video into consecutive frames. Then we select N consecutive frames from each video and encapsulate them as a volume to feed the 3D convolutional model. The 3D convolutional model is trained for outputting closer to the predefined category results. Finally, the well-trained model is used for global exam events prediction. The framework is showed in Fig. 1.

Our method adopts 3D convolutional neural network architecture which is developed from C3D model [11, 13]. 2D CNN convolves the spatial dimensions only, whereas 3D

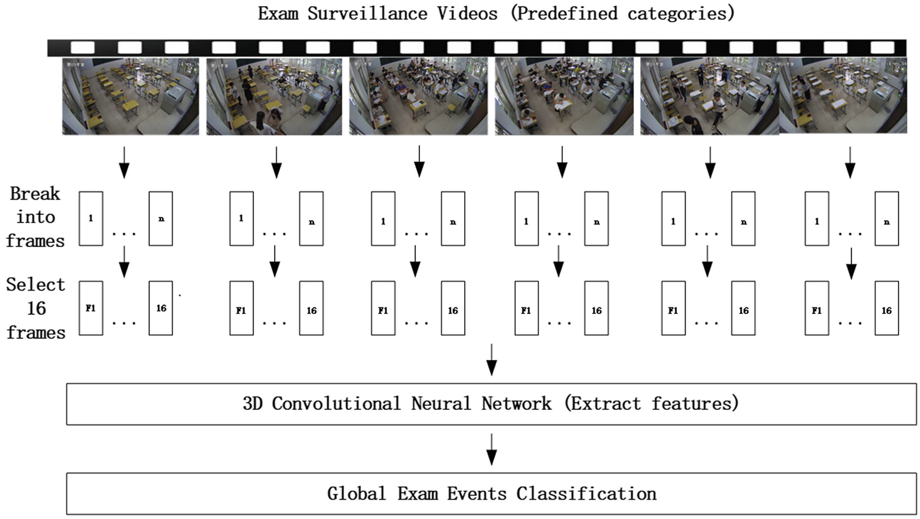

Exam Surveillance Videos (Predefined categories)
Break into frames
1
...
n
Select 16 frames
F1
16
3D Convolutional Neural Network (Extract features)
Global Exam Events Classification

Fig. 1. Framework diagram of detecting global exam events

CNN is different from 2D CNN by adding a temporal dimension to the convolutional kernel. Some behaviors in global exam events take time to finish. The temporal information is required in detecting the events. Therefore, 3D CNN should be considered to be more suitable for implementation on the global exam events scene. Assuming the input is the $F = (c, f, w, h)$ (c: channel, f: frame number, w: width, h: height), 2D CNN or multi-frames 2D CNN only uses the two-dimensional filters (k, k) (k: kernel width, k: kernel height), which results in two dimensions output. As the Fig. 2 shows, 3D CNN does the convolution operation by using filters whose dimension is (d, k, k) (d: kernel temporal depth, k: kernel width, k: kernel height) which convolves the volume (f, w, h) of each channel. Then, add each channel of RGB together to get the output of three-dimensional feature map. Since the 3D CNN kernel convolves the adjacent frames, it reserves temporal information. In this way, 3D CNN model can extract the temporal dimension features of frames, which plays key role in extracting motion features. When the chunks of consecutive frames feed to the 3D CNN architecture, the 3D convolution kernel can encapsulate both temporal and spatial information and output a feature volume map.

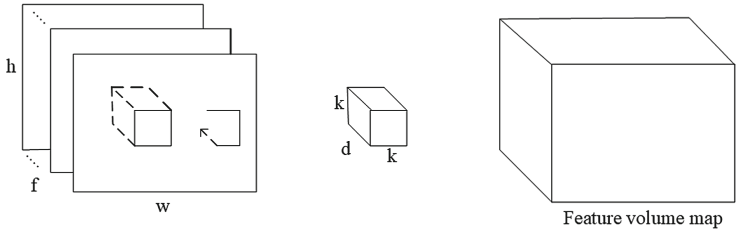


Fig. 2. 3D convolution

The experiment in [10] shows that gradually deeper network which gradually convolves and pools spatial and temporal information can achieve better results. The number of kernels also increase to generate different types of feature map. The pooling operation is max pooling, which reduces the feature map size except that the size of the temporal dimension is not changed in the first pooling layer. After convolution and pooling layers, it comes to fully connective layers to ensemble the features for the six categories of the global exam events. And the softmax function is used for normalizing the results and the cross entropy function is applied for minimizing the gradient loss.

Training: The 3D CNN is trained to extract spatio-temporal features from a given input which is a chunk of consecutive frames. The input dimension is denoted as,

$$x_i \in \mathbb{R}^{b \times f \times w \times h \times c}$$

where b stands for the sample number for a batch, f is the frame number in a chunk. w and h represent the width and height of each frame. c is channel number. Due to the RGB form, the initial input channel is 3.

After convolution and pooling layers, a feature map is produced. Its dimension is denoted as below, where p represents the pooling times, $NumF$ stands for the number of filters in the layer.

$$s_i \in \mathbb{R}^{b \times \frac{f}{p/2} \times \frac{w}{p} \times \frac{h}{p} \times NumF}$$

Our goal is to output vectors of $NumC$ categories, which represent for $NumC$ global exam events.

$$z_j \in \mathbb{R}^{b \times NumC}$$

We use the softmax function to normalize the output components corresponding to each category.

$$y_j = \frac{e^{z_j}}{\sum_{k=1}^{NumC} e^{z_k}} \tag{1}$$

The loss function uses cross entropy function. In order to minimize the gratitude between predefined label and forward propagation result, the parameters in the 3D CNN model is adjusted in the backward propagation phase,

$$loss\left(y'_j, y_j\right) = -\sum_{k=1}^{NumC} y'_j \log y_j \tag{2}$$

where y'_j represents the ground true label of the video frames.

3 Experiments and Evaluation

In the experiments, the input is 16 consecutive frames from each predefined video clips. By using the 3D convolution operation, the spatial-temporal information is extracted from the input volume and encapsulated in the output. After the convolutional layer, the pooling layer scales down the spatial sizes and merges the temporal size. Going through two fully connective layers, the output is mapped to six likely outcomes which are six categories of the exam events.

3.1 Dataset

Since there is no available public global exam events dataset, we establish our own dataset. The exam surveillance videos are collected from various exam rooms in primary or secondary schools. Some videos show different perspectives due to camera angles. All videos are in "avi" format with a frame rate of 25FPS. There are 916 videos in total. The global exam events are manually defined into six categories. They are "empty exam room status", "examinees entrance", "distributing papers", "on-exam status", "examinees departure", "collecting papers". Each category includes 153, 162, 148, 180, 113, 160 videos separately. 25% dataset is used for testing. Due to the uneven distribution of samples, the weight for each category will be considered in the following classification result. These six categories can generally classify different phases of exams. Each category is divided into 20 groups with 18 videos of the same behavior contained in one group. The length of each video is generally around 10 s. Some illustrations in each defined exam event category are showed in Fig. 3.

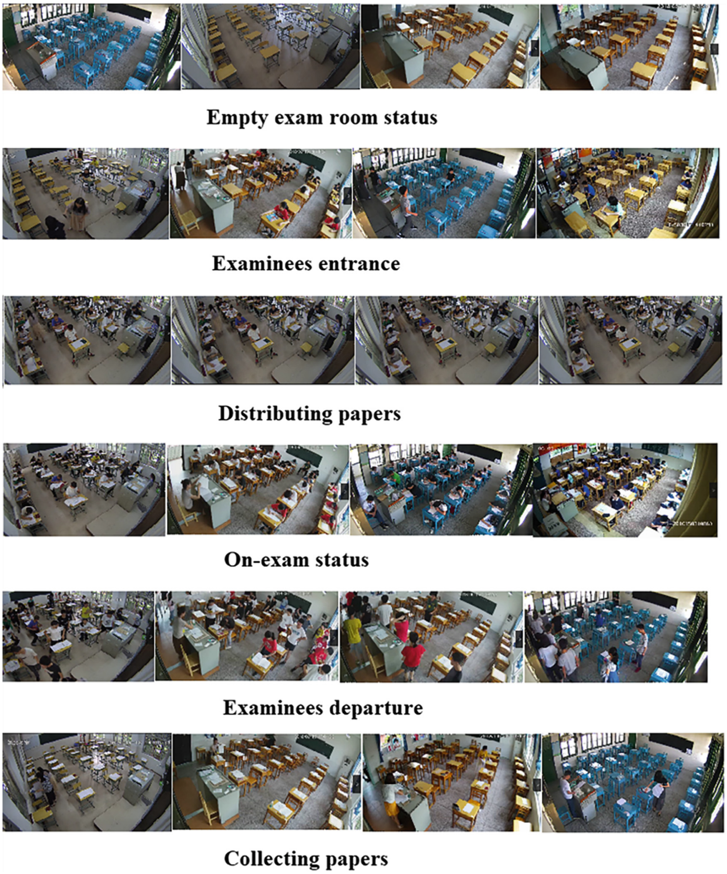
Empty Exam Room Status. Before the exam begins or when the exam ends, there is no one (students or invigilators) in the classroom. This event ends when someone open the door and enters the room.

Examinees Entrance. This event begins when examinees begin to enter the room from the front door. They receive security inspection by the invigilators and walk around the room to find their seat to sit down. This event ends when all the examinees have entered to room and sit down.

Distributing Papers. After all the examinees have entered the room and sit down, invigilators give out papers to examinees. Mostly, they walk to every examinee's seat and hand out the papers to the examinees one by one. Sometimes invigilators choose to give the papers to the first student of each row. This event ends when all papers are handed out to all the examinees.

On-exam Status. When exam begins, examinees begin to do the exam. Mostly, invigilators stay in front or back of the classroom and watch over the examinees. Sometimes, invigilators go around the classroom for inspection. This event ends when exam time is up, every examinee stops to do the exam.

Examinees Departure. After the on-exam status, examinees stand up and walk to the front door of the room for leaving from the room. This event ends when all the examinees leave and there are no examinees in the room.



Empty exam room status

Examinees entrance

Distributing papers

On-exam status

Examinees departure

Collecting papers

Fig. 3. Illustrations in six pre-defined global exam events. According to time sequence for an exam, there are phases of Empty exam room, Examinees entrance, Distributing papers, On-exam status, Examinees departure, Collecting papers, separately. These six global exam events depict the whole exam process.

Collecting Papers. After the examinees departure, opposite to distributing papers, invigilators come to each seat to collect the exam papers and sort them together. This event ends when there is no papers on each desk.

3.2 Parameters

We use Tensorflow [14] framework to implement our experiment. We firstly break the predefined videos to clips. Through iteration testing, the optimal value of consecutive frames is 16 which is enough for depicting a completed global exam event. The frames are cropped to 112 × 112 with channels of 3. The temporal dimension of the 3D kernel is set to 3 this experiment as it has been shown that 3 × 3 × 3 convolution kernel has the best performance [11]. The 16 consecutive frames are treated as a volume and each time we feed the model with 10 volume samples as a batch, thereby our input is of 5 dimensional tensors consisting of $10 \times 16 \times 112 \times 112 \times 3$ (batch-size, frames-per-clip, crop-size, crop-size, channels).

There are 5 convolutional layers with 8 times convolutional operations followed by ReLU activation function, 5 pooling layers, two fully connected layers, and a softmax output layer in total. All 3D convolution kernels are in size of $3 \times 3 \times 3$ with a stride of $1 \times 1 \times 1$. As the network goes deeper, the number of kernels increase from 64, 128, 256 to 512. With the purpose of preserving the early temporal information, the pool1 kernel size is $1 \times 2 \times 2$ with stride $1 \times 2 \times 2$, and all other pooling layers are $2 \times 2 \times 2$ with stride $2 \times 2 \times 2$ to scale down the spatial features and merge the temporal features. The fully connected layer output 4096 units which are then passed to the classification layer for classification. Finally, we get the classification label through the softmax layer. The 3D CNN architecture which we apply in our experiment is presented in Table 1.

Table 1. The architecture of the adopted 3D convolutional neural network. The architecture consists of 5 convolution layers, 5 pooling layers, 2 fully layers and 1 softmax output layer. Detailed descriptions are given in the text.

Conv1	Input:	[10, 16, 112, 112, 3]	Conv4b	Input:	[10, 4, 14, 14, 512]
	Output:	[10, 16, 112, 112, 64]		Output:	[10, 4, 14, 14, 512]
Pool1	Input:	[10, 16, 112, 112, 64]	Pool4	Input:	[10, 4, 14, 14, 512]
	Output:	[10, 16, 56, 56, 64]		Output:	[10, 2, 7, 7, 512]
Conv2	Input:	[10, 16, 56, 56, 64]	Conv5a	Input:	[10, 2, 7, 7, 512]
	Output:	[10, 16, 56, 56, 128]		Output:	[10, 2, 7, 7, 512]
Pool2	Input:	[10, 16, 56, 56, 128]	Conv5b	Input:	[10, 2, 7, 7, 512]
	Output:	[10, 8, 28, 28, 128]		Output:	[10, 2, 7, 7, 512]
Conv3a	Input:	[10, 8, 28, 28, 128]	Pool5	Input:	[10, 2, 7, 7, 512]
	Output:	[10, 8, 28, 28, 256]		Output:	[10, 1, 4, 4, 512]
Conv3b	Input:	[10, 8, 28, 28, 256]	FCNet1	Input:	[10, 1, 4, 4, 512]
	Output:	[10, 8, 28, 28, 256]		Output:	[10, 4096]
Pool3	Input:	[10, 8, 28, 28, 256]	FCNet2	Input:	[10, 4096]
	Output:	[10, 4, 14, 14, 256]		Output:	[10, 4096]
Conv4a	Input:	[10, 4, 14, 14, 256]	Out	Input:	[10,4096]
	Output:	[10, 4, 14, 14, 512]		Output:	[6, 10]

3.3 Results

Eventually, we get a six-category global exam event discrimination of 93.94% by using our method. Due to the pioneer of this work, there are few baselines able to be compared. To the best of our knowledge, our work is the first one to detect the global exam events and the approach in [15] is the most related work. Therefore, we choose [15] as our baseline to demonstrate the performance of our work. Table 2 compares the results using 3D CNN networks with two-classifier using HOG features [15] to demonstrate its superior performance.

Table 2. Comparation with other method

Method	Accuracy
3D CNN	93.94%
[15]	86.1%

Table 3 presents the performance of 3D CNN on classifying the six global exam events and compared with Depth-1 (2D CNN) method. It is worth mentioning that the data in Table 3 is weighted-calculated as a result of the uneven distribution samples. Obviously, empty exam room status event and on-exam status event have the best performance compared to other four events, whereas examinees departure event has the poorest accuracy. This may be due to the short period of time this event takes place, which lead to less training data of this event. A controlling experiment is also carried out by decreasing the temporal depth into 1, which means the whole 16 frames are convolved in 2D way separately. It is interesting to observe from Table 3 that the accuracy of three global exam events decreases when it comes to Depth-1 (2D CNN), they are distributing papers event, on-exam status event as well as collecting papers event. We believe it is due to that these three events are mostly completed through a period of time and are easily-confused from static images. For example, distributing papers is really similar with collecting papers except opposite directions. Reasonably, 3D CNN has higher accuracy on these three events due to its extraction of temporal features, which also demonstrate that 3D CNN functions better under global exam events scenes. Overall, 3D CNN outperforms Depth-1 (2D CNN) method as can be seen weighted average performance in Table 3.

Figure 4 presents the normalized confusion matrix results of the experiment. From the confusion matrix, we can find that the distributing papers event and the examinees departure event have relatively lower accuracy compared with other events, where distributing papers event has chance to be confused with on-exam status, examinees entrance and examinees departure event, examinees departure event tends to confuse with distributing paper event and collecting paper event. It may be caused by the lower discrimination of actions taken in these events and the weakness of this model to classify events when someone or several ones hold the standing position.

To conclude, the experiment result demonstrates that by capturing both spatial and temporal features simultaneously, our model has satisfactory performance for the global exam events recognition.

Table 3. Performance of the two methods. The performance is evaluated in terms of accuracy (Acc.), precision (Pre.), recall and F1 score. The highest value of each case is in bold. Due to the uneven distribution of samples, the weighted average, denoted as W-Mean, is calculated.

Events	3D CNN				Depth-1 (2D CNN)			
Measure	Acc.	Pre.	Recall	F1	Acc.	Pre.	Recall	F1
Empty-exam room	1	1	1	1	1	1	1	1
Examinees entrance	0.8966	**0.9630**	0.8966	**0.9286**	0.8966	0.8387	0.8966	0.8667
Distributing papers	**0.8723**	**0.9111**	**0.8723**	**0.8913**	0.8511	0.8696	0.8511	0.8603
On-exam status	**1**	0.9020	**1**	**0.9485**	0.9565	0.9362	0.9565	0.9462
Examinees departure	0.8462	0.9565	0.8462	0.8980	0.8462	0.9565	0.8462	0.8980
Collecting papers	**0.9714**	**0.9189**	**0.9714**	**0.9444**	0.9429	0.9167	0.9429	0.9296
W-Mean	**0.9394**	**0.9406**	**0.9394**	**0.9388**	0.9221	0.9230	0.9221	0.9220

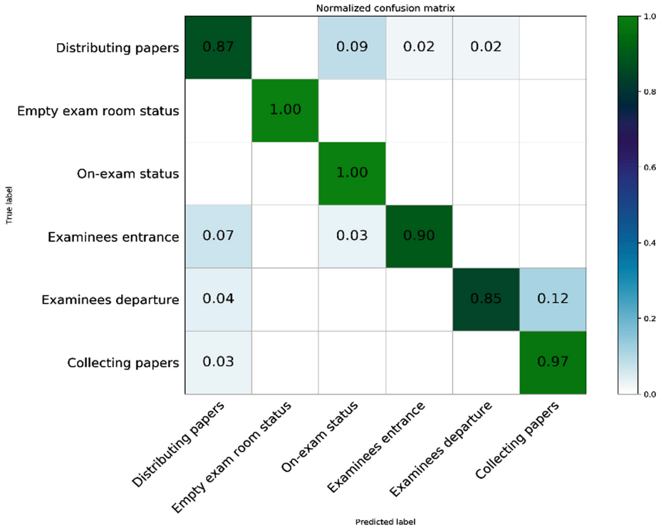


Fig. 4. Confusion matrix of six-category global exam events

4 Conclusions

This paper has proposed a structure of 3D convolutional neural network for the global exam event recognition and gains a promising accuracy result. It firstly built the examination video dataset which includes 916 surveillance videos of different classroom scenes and various camera angles. On the built dataset, it is done to test the proposed structure of 3D convolutional neural network to extract the spatial-temporal features from six kinds of exam events. The proposed algorithm achieved an accuracy of 93.94% to discriminate these six global exam events. Additionally, the superiority of 3D CNN model is evaluated by diminishing the depth kernel into 1.

More works could be explored in the future. For instance, when the accuracy of classifying the exam events is guaranteed, we should try shorter clips to improve its sensitivity and search for the boundary among these different exam events. Besides, examination video event detection is firstly implemented using 3D convolutional neural network, more advanced recognition technology needs to be explored and fused for better result.

Acknowledgement. This work is supported in part by National Natural Science Foundation of China under Grant 61802142, and Fundamental Research Funds for the Central Universities under Grant CCNU19QN031.

References

1. Adil, M., Simon, R., Khatri, S.K.: Automated invigilation system for detection of suspicious activities during examination. In: 2019 Amity International Conference on Artificial Intelligence (AICAI). IEEE (2019)
2. Cote, M., Jean, F., Albu, A.B., Capson, D.W.: Video summarization for remote invigilation of online exams. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1–9 (2016)
3. Laptev, I., Marszalek, M., Schmid, C., Rozenfield, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
4. Kläser, A., Marszalek, M., Schmid, C.: A spatio-temporal descriptor based on 3D-gradients. In: British Machine Vision Conference, pp. 995–1004 (2008)
5. Oneata, D., Verbeek, J.J., Schmid, C.: Action and event recognition with fisher vectors on a compact feature set. In: 2013 IEEE International Conference on Computer Vision, pp. 1817–1824 (2013)
6. Wang, H., Kläser, A., Schmid, C., Liu, C.: Action recognition by dense trajectories. In: CVPR 2011, pp. 3169–3176 (2011)
7. Peng, X., Zou, C., Qiao, Y., Peng, Q.: Action recognition with stacked fisher vectors. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 581–595. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_38
8. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Commun. ACM **60**, 84–90 (2012)
9. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1725–1732 (2014)
10. Feichtenhofer, C., Pinz, A., Zisserman, A.: Convolutional two-stream network fusion for video action recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1933–1941 (2016)
11. Tran, D., Bourdev, L.D., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3D convolutional networks. In: 2015 IEEE International Conference on Computer Vision (ICCV), pp. 4489–4497 (2015)
12. Wang, L., Li, W., Li, W., Gool, L.V.: Appearance-and-relation networks for video classification. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1430–1439 (2018)
13. Ji, S., Xu, W., Yang, M., Yu, K.: 3D convolutional neural networks for human action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **35**(1), 221–231 (2013)

14. Abadi, M., et al.: TensorFlow: large-scale machine learning on heterogeneous distributed systems. arXiv:abs/1603.04467 (2015)
15. Ding, M., Zhao, J., Hu, F.: Abnormal behavior analysis based on examination surveillance video. In: 2016 9th International Symposium on Computational Intelligence and Design (ISCID), 01, pp. 131–134 (2016)