


RESEARCH

Open Access



# Method for detection of unsafe actions in power field based on edge computing architecture

Yanfang Yin<sup>1</sup>, Jinjiao Lin<sup>2</sup>, Nongliang Sun<sup>1</sup>, Qigang Zhu<sup>1</sup>, Shuaishuai Zhang<sup>1</sup>, Yanjie Zhang<sup>3</sup> and Ming Liu<sup>1\*</sup> 

## Abstract

Due to the high risk factors in the electric power industry, the safety of power system can be improved by using the surveillance system to predict and warn the operators' nonstandard and unsafe actions in real time. In this paper, aiming at the real-time and accuracy requirements in video intelligent surveillance, a method based on edge computing architecture is proposed to judge unsafe actions of electric power operations in time. In this method, the service of unsafe actions judgment is deployed to the edge cloud, which improves the real-time performance. In order to identify the action being executed, the end-to-end action recognition model proposed in this paper uses the Temporal Convolutional Neural Network (TCN) to extract local temporal features and a Gate Recurrent Unit (GRU) layer to extract global temporal features, which increases the accuracy of action fragment recognition. The result of action recognition is combined with the result of equipment target recognition based on the yolov3 model, and the classification rule is used to determine whether the current action is safe. Experiments show that the proposed method has better real-time performance, and the proposed action cognition is verified on the MSRAction Dataset, which improves the recognition accuracy of action segments. At the same time, the judgment results of unsafe actions also prove the effectiveness of the proposed method.

**Keywords:** Unsafe action prediction, Edge computing, The Temporal Convolutional Neural Network (TCN), Gate Recurrent Unit (GRU)

## Introduction

Generally, power working is carried out in the environment of high voltage and large current, any non-standard actions may lead to major safety accidents. In order to ensure the safety of operation, the power industry has strict operation procedures and training to regulate the behavior of operators. However, in the actual operation process, there will still be problems such as operators not strictly following the safety operation procedures due to the negligence of the operators or the lack of awareness of the risk, which will raises great safety concerns. In order to strengthen the safety management and control of power operation sites, numerous important departments

or workplaces are equipped with monitoring systems [1]. Using this video information to timely discover, remind or stop the operator's possible violations will reduce the occurrence of relevant safety accidents.

However, the analysis of surveillance video still relies on manual work at present. This way of working has the following disadvantages: first, the detection of abnormal behaviour depends on the skill level of the viewer, and there are few experienced experts; second, long-term watching is easy to lead to missed inspection defects and missed correction of violations due to visual fatigue, leading to safety risks not found in time [2]. Therefore, making full use of the current artificial intelligence technology to realize real-time intelligent analysis and judgment of video information is an effective way to improve the safety

\*Correspondence: [skd992365@sdust.edu.cn](mailto:skd992365@sdust.edu.cn)

<sup>1</sup>Shandong University of Science and Technology, 266590 Qingdao, China  
Full list of author information is available at the end of the article

of power system equipment operation and reduce the incidence of safety accidents.

At present, the intelligent video monitoring system based on artificial intelligence technology mainly involves the research of foreign matter identification of transmission line, line fault, abnormal state analysis of substation, illegal invasion of personnel and so on [3–6]. However, there is little research on real-time prediction and alarm of nonstandard and unsafe actions based on the operation specification of power systems in specific environments, so as to reduce the probability of accidents.

Therefore, in view of the unsafe actions that violate the operation specifications, such as unlocking without five-prevention keys, electricity testing without insulated gloves, etc., a security action recognition architecture based on edge cloud architecture is proposed in this paper. Based on this architecture, an end-to-end unsafe action determination method is deployed to the edge, and the learning of the model is deployed to the cloud, providing more real-time and more accurate services. The architecture is shown in Fig. 1.

This study offers three contributions as follows:

1. In this paper, an unsafe action prediction architecture based on edge cloud technology is designed. In this architecture, the decision of unsafe actions is deployed to the edge, which improves the real-time performance of the decision. At the same time, the method of model relearning and synchronous updating is proposed to realize the continuous evolution of the model, which enhances the accuracy and reliability of the identification and improves the ability to identify new unsafe actions.
2. In order to improve the accuracy of action recognition, an action recognition model is proposed

in this paper, in which the TCN is used to extract local temporal features, and a cyclic neural network GRU is added to extract global temporal features. Experimental results show that the model has better performance than the single time series feature extraction method.

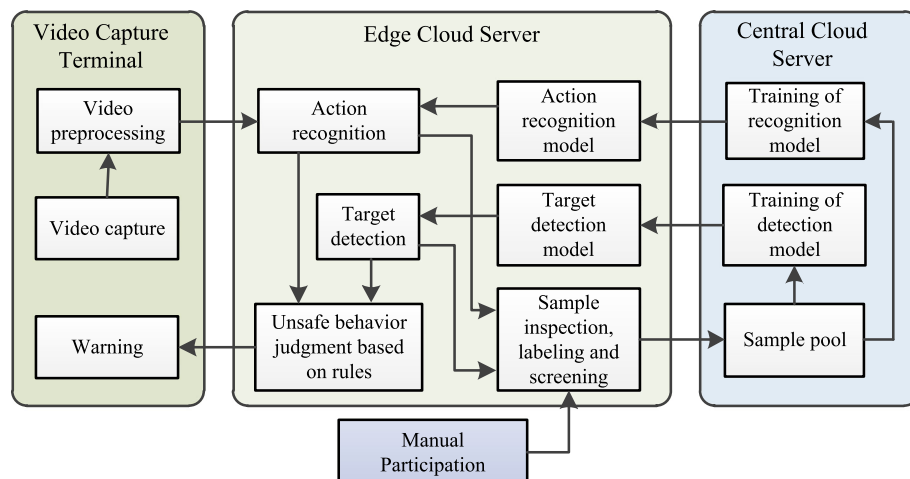
3. In this paper, a new method of judging unsafe action in electric power operation is proposed. The method combines the recognition results of action segments and target detection results, and determines unsafe actions based on rule classification.

## Related work

### Edge computing

The architecture design of edge computing dates back to 2009 when cloudlet [7] was proposed in Carnegie Mellon University. In 2011, Cisco first proposed the concept of fog computing [8]. It extends cloud computing by introducing an intermediate fog layer between mobile devices and cloud, which solves the problems of cloud computing's inability to perceive location and high latency. In 2016, the team [9] of Wayne State University gave the formal definition of edge computing for the first time and studied the application scenarios of edge computing. Then, with the joint release of edge computing reference architecture 3.0 [10] by ECC and AII in 2018, artificial intelligence solutions based on edge computing have become a research hotspot.

Zhao [11] proposed an intelligent edge computing technology based on federated learning, in which the edge development board realizes the identification of video monitoring, the enterprise server implements the model learning, and the cloud is responsible for combining the models of each enterprise and updating the model. Jia [12] discussed the application prospect of edge computing



**Fig. 1** Prediction framework of unsafe actions in power operation based on edge cloud

model based on distributed data collection and processing in intelligent video detection. In reference [13], the face recognition application is moved from the cloud to the edge, which greatly reduces the response time. In the paper [14], a method of constructing a face detection video monitoring system based on Mobile Edge Computing (MEC) is proposed. This method uses different detection algorithms in the edge and cloud, and determines whether it needs to be sent to the cloud according to the confidence of edge detection. Wang et al. [5] proposed a transmission line online monitoring system architecture based on ubiquitous Internet of Things by studying image recognition and mobile edge computing technology based on deep learning. Lu et al. [6] proposed a method based on edge calculation and deep learning for transmission line foreign matter detection. However, the existing edge detection model does not have the ability of relearning, and the model can not be improved, so it does not have the ability to identify new class.

### Action recognition

The traditional method of action recognition is to extract the temporal and spatial features of video images manually, and then classify the actions based on these features. Klaser et al. [15] extended the Histogram of Oriented Gradient (HOG) feature of static image to the space-time dimension, and proposed 3D HOG feature to represent action. Wang et al. [16] proposed the dense trajectories (DT) algorithm, which extracts and uses Histograms of Oriented Optical Flow (HOF), HOG and Motion Boundary Histograms (MBH) to represent actions. Ha [17] proposed a violence detection method in surveillance video system. The proposed method estimates the motion vector using the Combined Local-Global approach with Total Variation in the object region. The above method of action recognition uses the method of manually extracting the spatio-temporal characteristics of the image, which is inefficient, heavy workload and limited recognition ability.

Action recognition based on deep learning is an end-to-end method. This method uses the depth model to extract the spatial and temporal features contained in the video, and then classifies them. Simonyan K [18] first proposed action recognition method based on dual stream convolution network, which uses spatial stream network and time stream network to process spatial and temporal information in video separately. Feichtenhofer et al. [19] explored a variety of spatial and temporal information fusion methods on the basis of dual stream convolution network. The deep learning framework based on 3D convolution neural network can directly extract the temporal and spatial features of video [20–24]. The idea is to treat the video as a spatiotemporal cube, that is, the 2D convolution operation in the spatial domain is naturally extended to the 3D

convolution operation in the spatiotemporal domain by adding time dimension.

Recurrent Neural Networks (RNN), especially Long Short-Term Memory (LSTM) networks and GRU networks, have strong ability to extract temporal features [25–27]. In reference [28–30], a 2D Convolutional Neural Network (CNN) is used to extract spatial features of video images, and LSTM or GRU, a variant of RNN, is used to extract temporal features of actions, and good recognition rate is obtained. In order to simulate the spatiotemporal evolution of different actions and extract different spatiotemporal features, Song et al. [31] constructed an action recognition model based on CNN using LSTM. The model used different levels of attention to learn the recognizable joints of bones in each input frame. Tae Soo Kim et al. [32] propose to use a new class of models known as TCN for 3D human action recognition. TCN provides a way to explicitly learn readily interpretable spatio-temporal representations for 3D human action recognition. Methods used in the above literatures can only classify the action after it is completed, and can't predict the future action, so they can't be used for action early warning.

### Action prediction

In view of the early warning function, the system should not only identify the action, but also predict the future action. That is to say, the system should have the ability to classify actions according to the segments of some actions. Fragkiadaki et al. [33] proposed an Encoder-Recurrent-Decoder model based on recurrent neural network for recognition and prediction of human body pose in videos and motion capture. Jain et al. [34] introduced RNN Structure, which combined high-order spatiotemporal images and RNN to predict actions in a short time. Martinez et al. [35] modified the standard RNN model for human motion to form a simple and extensible RNN architecture, and good performance of human motion prediction was obtained. Keet et al. [36] proposed a new Latent Global Network based on adversarial learning for action prediction. In this model, skeleton is used for action prediction, which aims to identify an action from partial skeleton sequence with incomplete action information. Kong et al. [37] adopted an adversarial learning scheme to learn action features, extract model parameters and classifier parameters, and generate optimized features for action prediction. The method has achieved good prediction accuracy. RNN has a strong ability to extract temporal features, but the accuracy of RNN (LSTM, GRU, etc.) for long sequences is low. In order to improve the accuracy of motion prediction and obtain better performance, TCN is used to extract local temporal features, and recurrent neural network GRU is used to extract global temporal features.

### Action detection architecture based on edge cloud

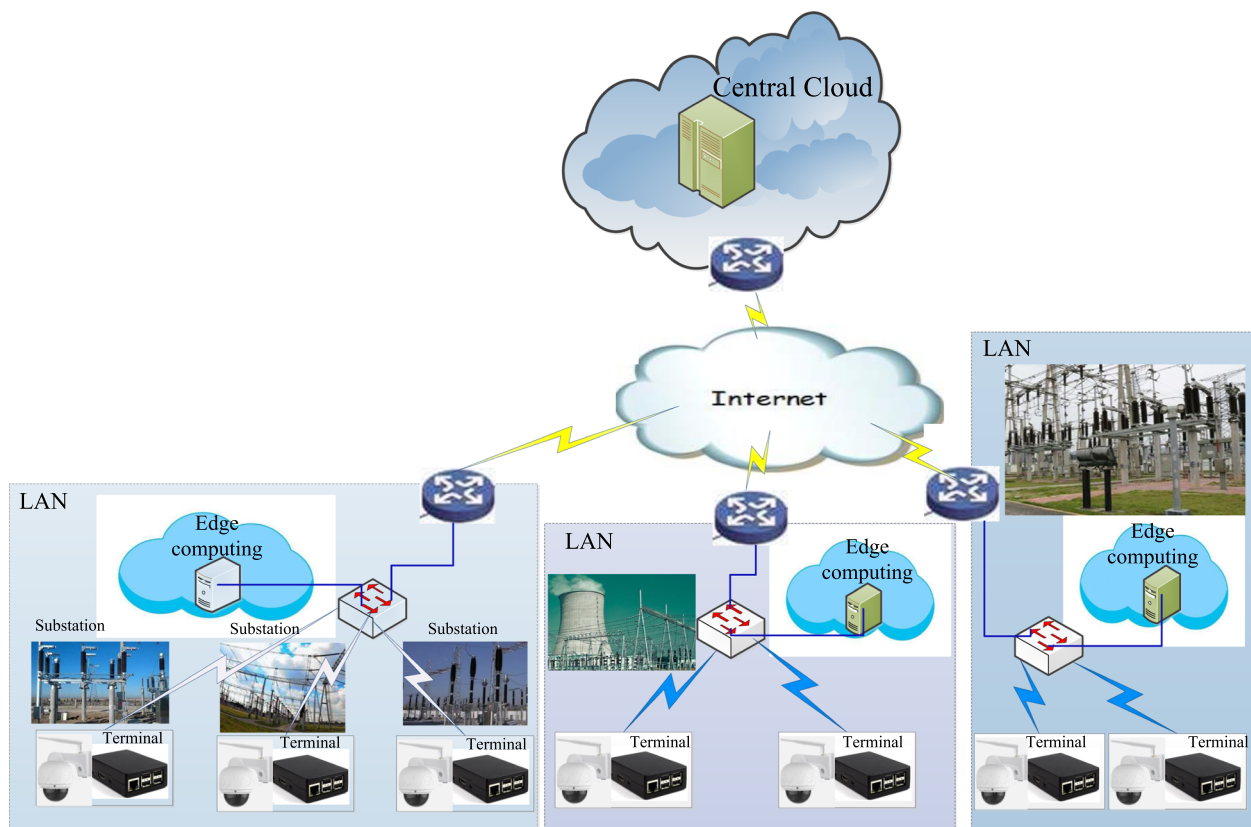
In order to judge unsafe actions based on video information, deep network model is needed to complete action recognition and target detection. This scheme involves a large amount of data and requires high computational performance and real-time information processing ability. Video capture terminal is a common embedded system with limited computing power, so it will take a lot of time to complete model calculation and information processing. If the action is determined by the cloud, the transmission of a large number of video information on the Internet will take a certain amount of time, and there may be network congestion, so it is difficult to guarantee the real-time performance of information processing. However, the system for unsafe action recognition and early warning requires high real-time performance. Therefore, in this paper, an unsafe action decision framework based on edge cloud is designed. The overall network architecture is shown in Fig. 2.

The video capture terminal and the edge cloud are interconnected by the Local Area Network (LAN), and the edge cloud is connected with the central cloud through the Internet. The decision service of unsafe actions is

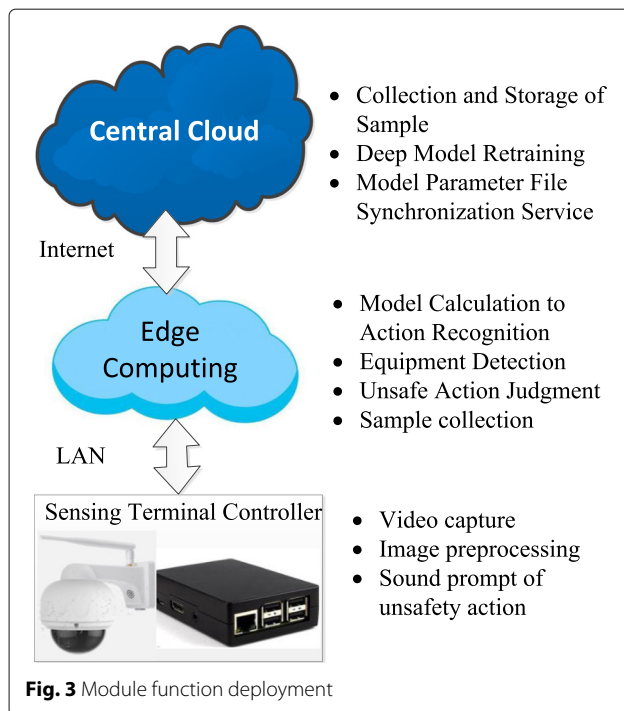
deployed to the edge to improve the real-time performance of identification. The model training module is deployed in the cloud. At the the same time, the cloud is also responsible for sample collection, module relearning and updating services to ensure the continuous learning and evolution of the model. The deployment of each functional module of the system is shown in Fig. 3.

The video image acquisition and simple preprocessing part of video image are deployed to the perception terminal. When the continuous image frames with human motion are detected by the perception terminal, these images will be standardized and scaled, and the data will be sent to the local edge computing entity using the Message Queuing Telemetry Transport (MQTT) protocol.

The decision service of unsafe action is deployed to local edge computing. When the unsafe action judgment entity at the edge receives the video image data sent by the terminal, it is handed over to the action recognition model for action recognition, and the data is sent to Yolo model for equipment target detection. Based on the identification results and detection results, the unsafe



**Fig. 2** Network architecture



action is judged by rule classification, and the final result is returned to the perception terminal through MQTT protocol. The human-computer interaction entity is used to collect, screen and label the newly discovered or newly defined unsafe action videos and new equipment test samples in a period of time, and upload them to the cloud sample pool by using the HyperText Transfer Protocol (HTTP).

The training of the model is deployed in the cloud. After receiving new samples, the cloud saves them to the sample pool through cloud storage services. When a certain number of new samples are reached, model learning is triggered. The new model files obtained by model re-learning are pushed to the edge through file synchronization service and MQTT protocol to achieve synchronous updating of cloud to edge models. The working principle framework of the whole system is shown in Fig. 4.

### Judgment of unsafe actions in power workplace

The power industry has strict regulations for the operators to work in the power field. For example, a person entering the power working place must wear safety helmets, the operators must wear insulating gloves for electricity testing, and the five-prevention keys must be used for unlocking. Otherwise, it will be unsafe actions. Therefore, the judgment of unsafe actions not only needs to identify the operator's action, but also to detect the necessary equipment for their operation.

### Power operation action recognition model based on TCN+GRU temporal feature extraction

A video of an operation action is a group of image sequences with temporal relationship, which contains not only the spatial characteristics of the action such as posture and the position relationship with other objects, but also the temporal information of the action, such as the change process of posture. According to the related work, RNN, especially LSTM network and GRU network, has a strong ability to extract temporal features. However, RNN network can not carry out parallel operation, and the number of operation steps should not be too large, otherwise it will take too long. TCN network is used to extract the temporal features of actions, which solves the problem of parallel computing in temporal feature extraction [38, 39]. However, the extracted features of TCN network also contain temporal feature information. In order to make full use of the advantages of RNN and TCN, an end to end unsafe action recognition model based on TCN+GRU for temporal feature extraction is presented in this paper. In this model, multi-layer TCN is used to extract local temporal features and down sample, which shortens the length of the time series. Then, the global temporal feature is extracted using GRU, which avoids the problem of long steps in GRU. Resnet50 network structure is used to extract spatial features. The overall structure of the action recognition model is shown in Fig. 5.

### Spatial feature extraction based on Restnet50

The action video of power operation is essentially a sequence of images, which contains the spatial information of the action, such as the shape of the limbs and the position relationship with other objects. In general, deep convolution network is used to extract image features. However, with the increase of depth, the ability of network representation is enhanced, and the difficulty of training is also increasing. In the process of back-propagation, for the deep neural network, the continuous cumulative multiplication can easily lead to the gradient too small or too large, resulting in the degradation of the learning ability of the deep network.

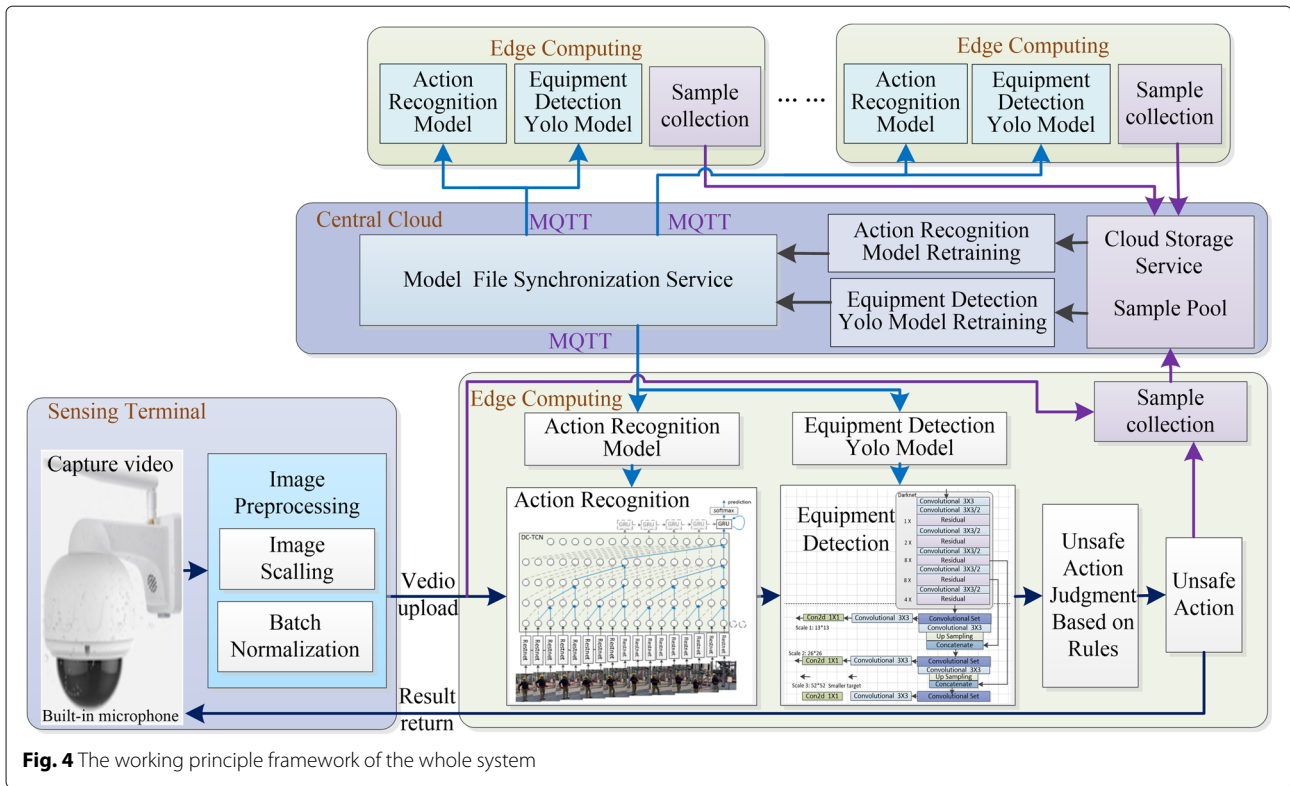
To solve this problem, the residual structure network Restnet50 presented in reference [40] is used to extract the spatial features of video actions. The Restnet50 network consists of several residual blocks. The output of a residual block is expressed as:

$$H(x) = F(x) + x \quad (1)$$

Where  $x$  is the input of the residual block,  $F(x)$  is the output of the network in the block, and  $H(x)$  is the output of the residual block.

Therefore, during forward propagation, the features of the upper layer can be reused in the next layer. At





**Fig. 4** The working principle framework of the whole system

the same time, when back propagating, the gradient is expressed as:

$$\nabla x = \nabla H(x) \nabla F(x) + \nabla H(x) \quad (2)$$

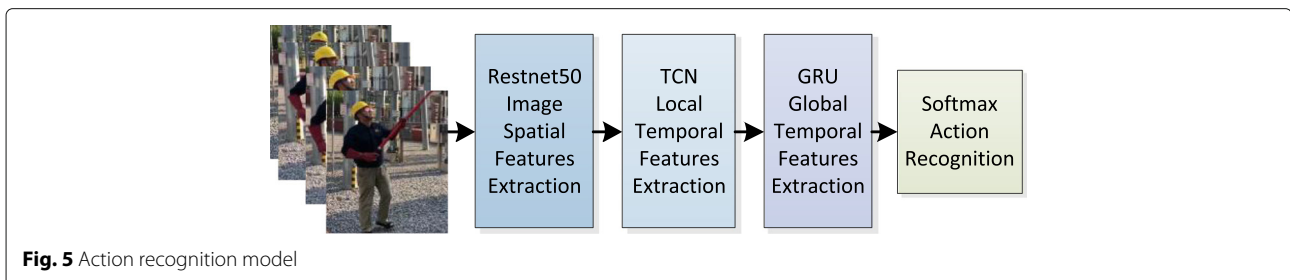
This ensures that when  $\nabla H(x) \nabla F(x) \rightarrow 0$ , the gradient of the lower layer  $\nabla H(x)$  can be directly transferred to the upper layer for updating, thus solving the problem of learning ability degradation.

Restnet50 uses residual blocks repeatedly to extract features, including 16 residual blocks and 50 layers in total. Since Restnet50 has many layers and many parameters, it is difficult to learn. Therefore, drawing lessons from the idea of transfer learning, based on the Restnet50 pre-training model provided by Keas, the trained Restnet50 model is obtained by relearning from our sample set and synchronized to the edge computing to extract the spatial features of video.

#### Local temporal feature extraction based on TCN

TCN is a one-dimensional convolution in the time dimension. The characteristics of its convolution calculation determine that TCN has a very good local feature extraction ability. At the same time, because the discrimination of unsafe actions is mainly for the purpose of action prediction, only the frame information in front of the moment is used, and the information behind the moment can not be used. Considering the above problems, in this paper, TCN based on causal convolution is used to extract local temporal features in video.

After extracting spatial features by Restnet50 model introduced in the former section, the video spatial feature extraction sequence is set as:  $X = [x_1, x_2, \dots, x_t, \dots, x_T]$ , where  $x_t$  is the feature of frame image  $img_t$  extracted by Restnet50 model,  $X$  is the input sequence of the expanded TCN. Let the convolution kernel  $F = [f_1, f_2, \dots, f_K]$ , and



**Fig. 5** Action recognition model

the output of convolution after causal TCN convolution is  $Y = [y_1, y_2, \dots, y_T]$ , then:

$$y_t = (Y * X)(x_t) = \sum_{k=1}^K f_k x_{t-(K-k)} \quad (3)$$

In order to better extract local features and increase receptive field, this model uses two layers of  $k = 4$  TCN for convolution, and the output of each layer is pooled with step size of 2.

#### Global temporal feature extraction based on GRU

By TCN convolution, some temporal features have been extracted, but the output of TCN convolution is still a sequence, including temporal information. Since the cyclic neural network has the ability to describe the sequence information, it can often get more accurate results. Therefore, in this paper, GRU is used to further extract global features from the results of TCN convolution.

GRU has set up update gate  $z_t$  and reset gate  $r_t$ , the calculations for them shown as formula (4) and (5). The reset gate  $r_t$  determines how much information of the previous state  $h_{t-1}$  is used to calculate the candidate state  $\tilde{h}_t$  as shown in formula (6). The update gate  $z_t$  determines the amount of information that the current state  $h_t$  needs to obtain from the previous state  $h_{t-1}$  and candidate state  $\tilde{h}_t$  as shown in formula (7). The calculation formula is as follows:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (4)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (5)$$

$$\tilde{h}_t = \tanh(W \cdot [r_t * h_{t-1}, x_t]) \quad (6)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (7)$$

Where  $\sigma$  is the sigmoid function, by which the data can be transformed into a value in the range of  $0 \sim 1$  to act as a gate control signal.  $x_t$  is the input of the current time,

$h_{t-1}$  is the state of the previous time and  $h_t$  is the state of the current time.  $W_z$ ,  $W_r$  and  $W$  are the network parameters for calculating  $z_t$ ,  $r_t$  and  $\tilde{h}_t$ . The output of the last step of the GRU network contains all the features of the video. Therefore, the features output in the last step can be directly classified into two categories after being fully connected to predict unsafe actions.

#### End to end action recognition model

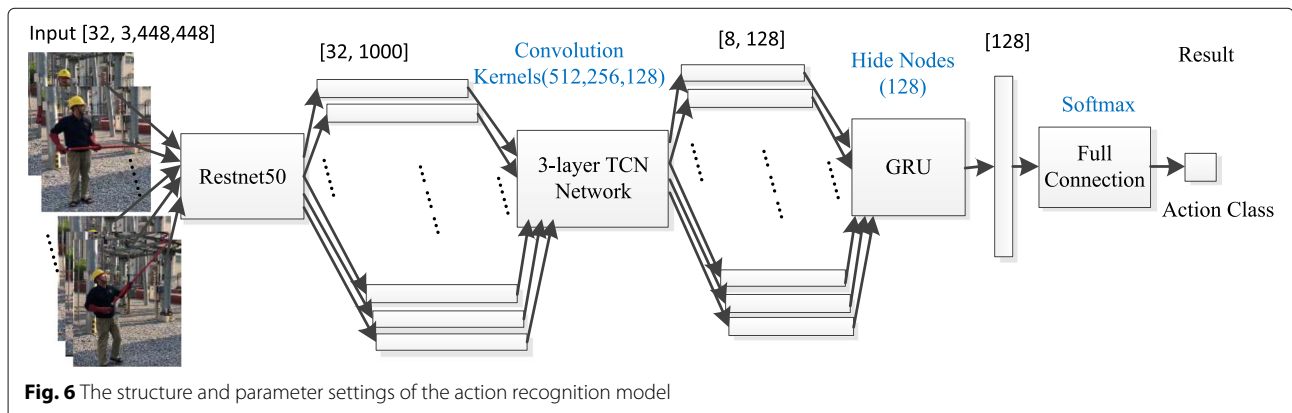
The structure and parameter setting of the end-to-end action recognition depth model proposed in this paper is shown in Fig. 6. After scaling and standardization pre-processing, a power operation action video is taken as the input of the action recognition model. It is expressed as:  $X = [img_1, img_2, \dots, img_T]$ , in this paper  $T=32$ , so the shape of  $X$  is  $[32, 3, 448, 448]$ .

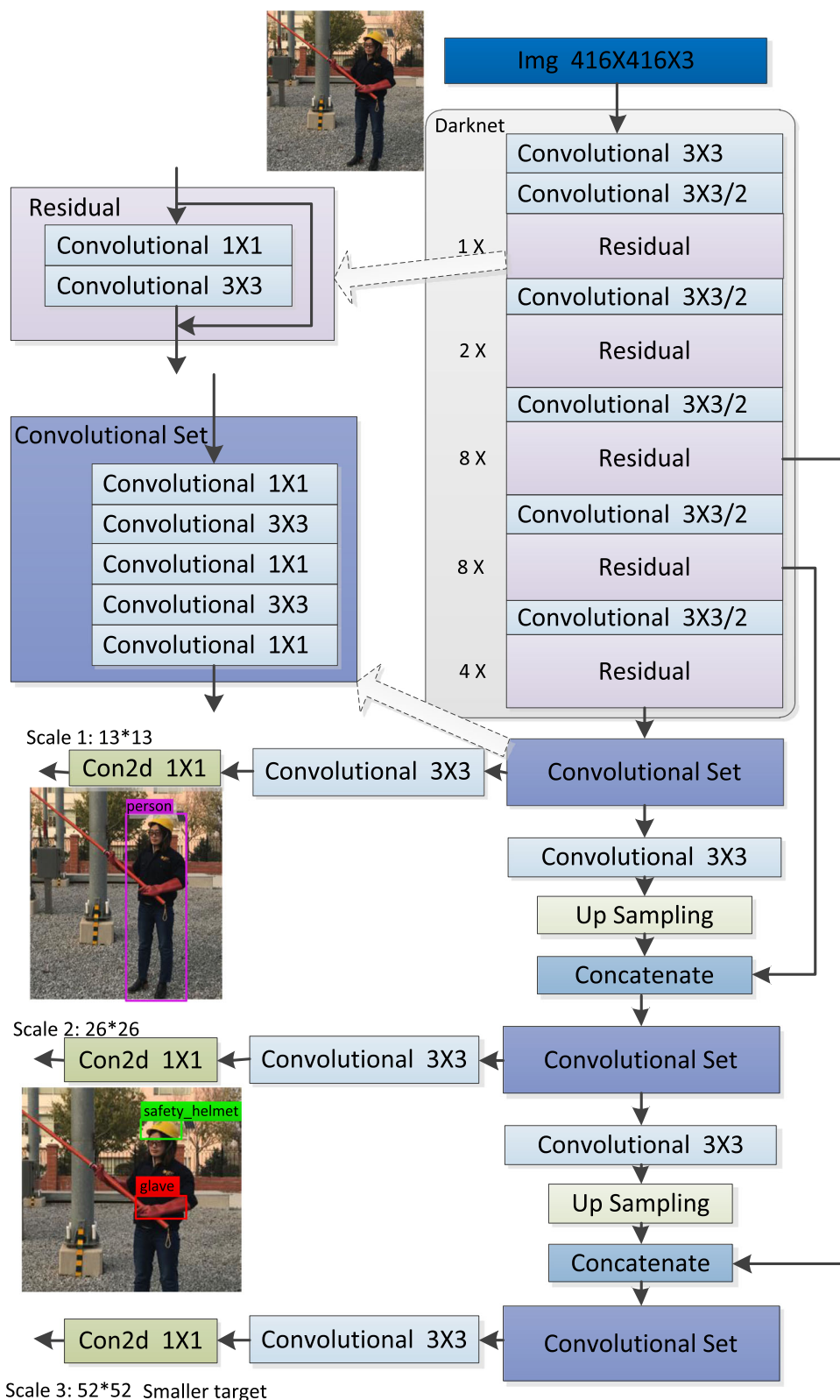
For each image  $[3, 448, 448]$ , the spatial feature is extracted by Resnet50 network, and the feature output of each image is 1000 dimensions. Therefore, after the input  $X$  passes through the Resnet network, the output shape is  $[32, 1000]$ . The 3-layer TCN network is used to extract local temporal features. The number of convolution kernels are set to (512, 256, 128). After extracting the local temporal features through the TCN network, the output shape is  $[8, 128]$ . The number of GRU hidden nodes is set to 128, and the dimension is [128] after global temporal features are extracted by GRU. By the softmax classifier, the class of action is output.

#### Equipment target detection based on Yolov3 model

The judgment of unsafe action is closely related to the necessary equipment. Therefore, equipment target detection is an important content of unsafe action judgment.

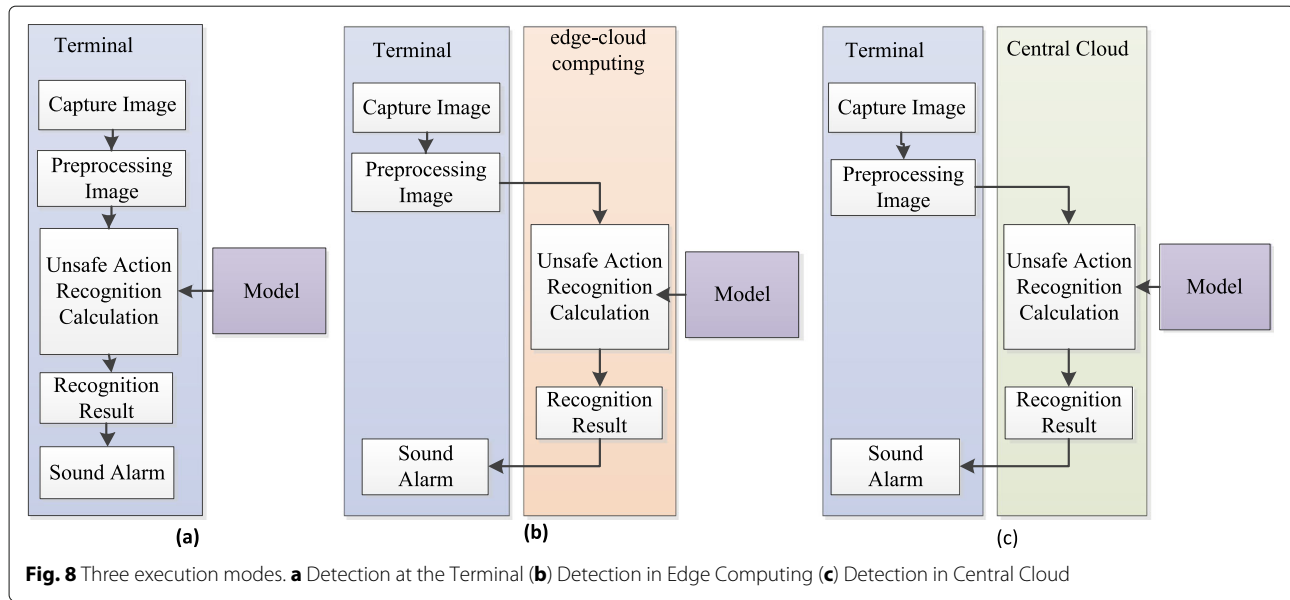
For object detection, Yolo model has achieved good results, which can quickly detect the target and mark the target position [41]. In this paper, yolov3 model [42] is used for real-time detection of necessary equipment such as safety helmet, insulating gloves, five-prevention





**Fig. 7** Yolov3 equipment detection model model





keys, etc. The structure of necessary equipment detection model based on Yolov3 model shown as in Fig. 7.

For the Darknet53 part of yolov3 model, this paper adopts the method of transfer learning, and retrain the model with our own labeled dataset. And ten kinds of equipment or objects, such as person, helmets, heads, insulated gloves, insulated poles, five-prevention keys and electric boxes, are mainly detected in this system. At present, the system can detect ten kinds of targets, such as person, helmet, head, insulated gloves, insulated pole, five-prevention and so on.

### Unsafe action judgment based on rule classification

The judgment rules of unsafe actions are defined based on the operation rules of power industry. For example, in view of not wearing safety gloves and unlocking without five-prevention keys, a group of unsafe action judgment rules can be defined as follows:

$$\begin{aligned}
 R1 : IF \text{ action} = \text{"electrictesting"} \wedge \text{"insulatedgloves"}\text{target} \\
 &= \text{false THEN act} = \text{unsafe} \\
 R1 : IF \text{ action} = \text{"unlocking"} \wedge \text{"fivepreventionkeys"}\text{target} \\
 &= \text{false THEN act} = \text{unsafe}
 \end{aligned}
 \tag{8}$$

Here, we define only two unsafe actions. If new unsafe actions need to be defined, rules can be added. Based on the output of action recognition model and Yolo detection model, unsafe actions can be identified by using the rule classifier.

### Experiment

In order to verify the real-time performance of edge computing Architecture, we build a proof of concept platform. The sensing terminal uses the Raspberry Pi 3B as the controller, and it is connected to the camera, which is a camera with a built-in speaker with a focal length of 12mm and a resolution of 1080p zoom. It's operating system is Raspbian, and it connects to the LAN network by Wi-Fi. The edge computing server CPU is Intel Xeon E5-2600 v3, the graphics card is NVIDIA Quadro K4200, and it's operating system is Ubuntu Linux. The terminal and the edge computing server are in the same local area network. The central cloud uses Alibaba cloud enterprise level universal ecs.g6.2xlarge, and its operating system is Ubuntu Linux. The edge computing connects to the central cloud through the Internet.

### Time validation of the architecture

After the action recognition model and Yolov3 model are trained, the entity of unsafe action detection ran in the

**Table 1** Time of different modes

| Mode                        | Transmission time (s) | Calculation time (s) | Total time (s) |
|-----------------------------|-----------------------|----------------------|----------------|
| Detection at the Terminal   | 0                     | 17.305 1             | 7.305          |
| Detection in Edge Computing | 0.390                 | 1.619                | 2.009          |
| Detection in Central Cloud  | 3.386                 | 5.899                | 9.285773       |

**Table 2** Number of samples of different length action fragments generated by MSR Data set

| The Length of action fragments | Number of training samples | Number of testing sample |
|--------------------------------|----------------------------|--------------------------|
| K=16                           | 10895                      | 3401                     |
| K=24                           | 7408                       | 2449                     |
| K=32                           | 4304                       | 1549                     |
| K=40                           | 2165                       | 860                      |
| K=48                           | 1157                       | 431                      |
| K=56                           | 803                        | 219                      |
| K=64                           | 713                        | 140                      |

terminal, the edge computing and the cloud in the three ways as shown in Fig. 8.

The transmission time  $T_{tr}$ , the calculation time of unsafe action judgment  $T_c$  and the total time  $T_{tol}$  are shown in the following formulas.

$$T_{tr} = (t_{video\_r} - t_{video\_s}) + (t_{out\_r} - t_{out\_s}) \quad (9)$$

$$T_c = t_{out\_s} - t_{video\_r} \quad (10)$$

$$T_{tol} = t_{out\_r} - t_{video\_s} \quad (11)$$

Where,  $video\_s$  is the time when the terminal sends video,  $t_{video\_r}$  is the time when the edge end or cloud receives the video,  $t_{out\_s}$  is the time when the identification service completes sending the identification result,  $t_{out\_r}$  is the time when the terminal receives the identification result.

In the three ways shown in Fig. 8, the time spent to determine whether a video action segment ( $K = 32$ ) is unsafe is shown in Table 1.

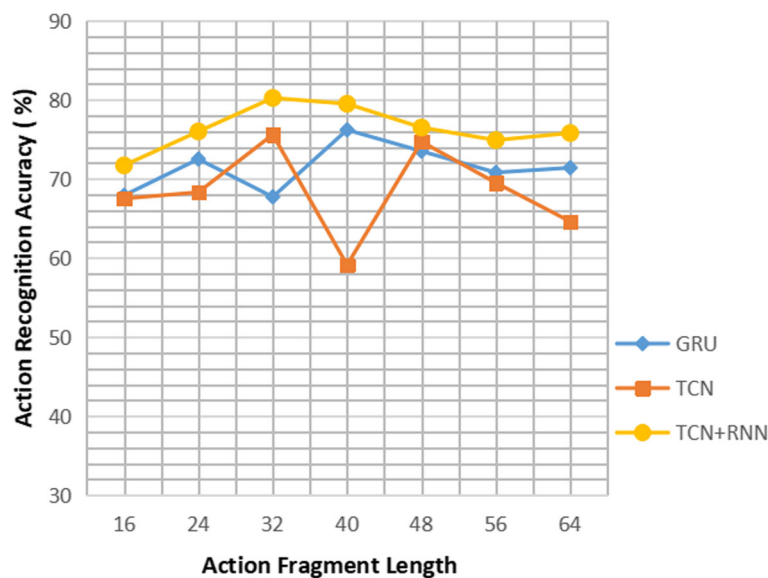
In the experiment, the recognition time of cloud is relatively high, which is due to the low performance of the cloud server we applied for. However, it can be seen from the table that the transmission time of the edge computing is far lower than that of the cloud, and the recognition calculation time is far lower than that of the terminal. The determination result can be returned in about 2 seconds using the edge computing architecture, which can meet the requirements of real-time performance.

### Verification of accuracy

In order to verify the effectiveness of the action recognition model based on TCN+GRU temporal features extraction, we use MSR action 3D standard dataset [43] to verify. The MSR action dataset recorded 20 kinds of actions, 10 subjects, and each object performed each action 2-3 times. The resolution of depth map is  $640 * 240$ . As in reference [38], 116 sample actions of 3 and 9 subjects are taken as testing set, and 451 sample actions of other subjects are used as training set.

Because the unsafe action judgment is not for a complete action, but for a part of an action video. Therefore, in order to verify the recognition ability of our method for action segment, we use the sliding window to obtain part of the action frame to represent an action segment. The number of samples obtained by using different sliding window length  $K$  is shown in Table 2.

In order to verify the superiority of TCN+GRU temporal feature extraction proposed in this paper, after extracting the spatial features, TCN [38], GRU [30] and TCN+GRU are used to extract temporal feature, and then the action

**Fig. 9** The recognition accuracy of action segments with different lengths in different models



**Fig. 10** Safe and unsafe actions

classification is carried out through softmax layer. To ensure fairness, the network model uses the same hyper parameters as shown in Fig. 5. The experimental results are shown as Fig. 9.

As shown in the experiment, we can find that no matter how much  $K$  is, the accuracy of action recognition based on TCN+GRU temporal feature extraction method is higher than the other two methods. The main reason is that the TCN + GRU temporal feature extraction method not only extracts global features but also local features, so it is less affected by the sample quality and better overcomes the shortcomings of single feature extraction method. This proves the superiority and rationality of the method proposed in this paper.

#### Detection of unsafe action in electric power operation

There are safety regulations for operation in the power plant, such as electric testing, unlocking, opening or closing switch etc.. For example, if you will perform electric testing, you must wear a safety helmet and gloves, otherwise it is an unsafe. In this experiment, we test two kinds of unsafe actions as shown in Fig. 10, one is electric testing without insulating gloves, and the other is unlocking without using five-prevention keys.

#### Video sampling

In the experiment, we asked four person to finish 25 groups of safe and unsafe actions as shown in Fig. 10, and captured images at the speed of 30 frames per second. And we obtained 100 video samples. The action types of the video samples are shown in Table 3.

#### Action recognition

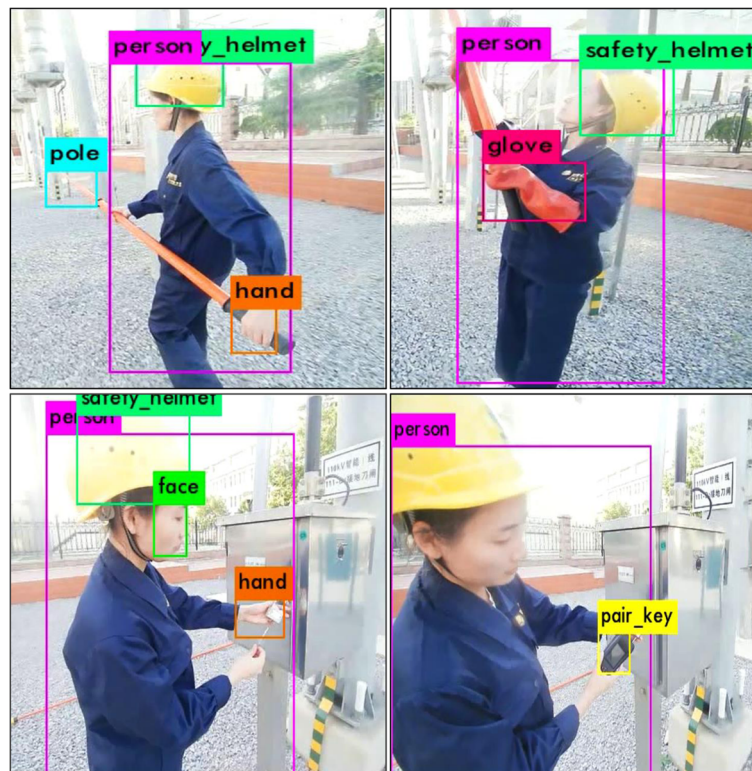
Because the action prediction is not to recognize the whole action, but to recognize the specific action by using a small part of the action frame. Therefore, for the processing of action video, one image is taken every three frames, that is, 10 images are taken from one second video. At the same time, the sliding window  $k = 32$  frames is used to collect action fragments, so the duration of one action segment is 3.2 seconds. Because some action videos take a long time, a total of 887 sets of motion clips are obtained. 100 groups of action fragments are randomly selected as the verification set and the rest as the test set. Because the distinction between the two types of actions is obvious, the recognition rate of the two types of actions is 100% by using the proposed action recognition method.

#### Object detection

In order to detect objects with yolov3 model, we label 556 images with labeling tool. The detection objects include person, helmet, face, insulated gloves, pair key, insulating pole, hand and so on. The original yolov3 model is retrained by using the marked pictures to get a new model. The results of some object recognition are shown in Fig. 11.

**Table 3** The action types of the video samples

| Actions             | Safe actions | Unsafe actions |
|---------------------|--------------|----------------|
| Electricity testing | 28           | 32             |
| Unlocking           | 20           | 20             |



**Fig. 11** Object detection results

### Unsafe action judgment

Yolov3 model is not stable for target detection, and it can not be detected in every frame. Therefore, in order to improve the reliability of target detection, we use youloov3 to detect 32 frames of an action segment. As long as the equipment is detected in two images, the equipment is considered to be true, and then the unsafe action is determined by rules. The key will not be detected due to occlusion and other reasons, and the safety action may be judged as unsafe action, resulting in the accuracy rate of unsafe action detection reduced to 91% in the test set.

### Conclusion

In this paper, the problem of real-time judgment of unsafe actions in power operation is discussed, and the intelligent monitoring architecture based on edge cloud technology and the problem of judging unsafe actions are explored. According to the above analysis and experiments, it can be seen that:

1. The unsafe action detection architecture based on edge cloud architecture in this paper can solve the problem of network transmission delay and meet the needs of continuous learning and upgrading of the model.

2. By adding a GRU layer to extract the global temporal information, the proposed action recognition model increases the recognition ability of action segments. At the same time, combined with the results of equipment detection by Yolov3 model, the unsafe action can be judged by rule classification, and the purpose of early warning is achieved.

Due to the time limit, there are still some improvements in this paper, such as how to better combine the action recognition model and Yolov3 model, and more work needs to be done in the future.

### Abbreviations

TCN: The temporal convolutional neural network; GRU: A gate recurrent unit; RNN: Recurrent neural network; MEC: Mobile edge computing; HOG: The histogram of oriented gradient; HOF: The histograms of oriented optical flow; MBH: Motion boundary histograms; LAN: The local area network; MQTT: The message queuing telemetry Transport; HTTP: The hyperText transfer protocol

### Acknowledgements

Benjamin Millar, from the University of Tasmania in Australia, checked and revised the English expression and grammar of our paper.

### Authors' contributions

All authors have participated in conception and design, or analysis and interpretation of the data, drafting the article or revising it critically for



important intellectual content. The authors read and approved the final manuscript.

### Funding

This work was supported in part by the National Natural Science Foundation of China under Grant 61703243, in part by the Major Program of Shandong Province Natural Science Foundation under Grant ZR2018ZC0437

### Availability of data and materials

Since dataset and code involve our interests, we're sorry but we can't publish dataset at present. However, they are available on reasonable request.

### Competing interests

The authors declare that they have no competing interests.

### Author details

<sup>1</sup>Shandong University of Science and Technology, 266590 Qingdao, China.

<sup>2</sup>Shandong University of Finance and Economics, 250014 Jinan, China.

<sup>3</sup>Shandong Electric Power College, 250002 Jinan, China.

Received: 31 August 2020 Accepted: 1 February 2021

Published online: 22 February 2021

### References

- Ming L, Lin Y, Xianwei L, Fan Z, Jiaming Z (2018) Application of intelligent video analysis technology in power equipment monitoring. *Northeast Power Technol* 39(10):26–29
- Junhuang Z, Tingcheng H, Xiaoyu X, Wenjun F, Tingting Y, Yongjun Z (2020) Application of video image intelligent recognition technology in power transmission and transformation system. *China Electr Power* 54(11):1–12
- Qing L, Rongrong S, Huanbin C (2019) Application and practice of high voltage power line online video monitoring system. *Digit Technol Appl* 37(12):162–163
- Hui W, Mingjun L, Yingyi Y, Hao W, Qiang S (2019) Target detection method and Optimization for Substation Video Monitoring Terminal. *Guangdong Electr Power* 32(09):62–68
- Yanru W, Haifeng L, Lin L, Lanfang L, Zhenya Y (2019) Application of image recognition technology based on edge intelligent analysis in transmission line online monitoring. *Power Inf Commun Technol* 17(07):35–40
- Yanqiao L, Cuiying S, Hongwei C, Hongwei Y (2020) Foreign matter detection method for transmission equipment based on edge calculation and deep learning [J]. *China Electr Power* 53(06):27–33
- Satyanarayanan M, Bahl P, Caceres R, et al. (2009) The case for VM-based cloudlets in mobile computing. *IEEE Pervasive Comput* 8(4):14–23
- Bonomi F (2011) Connected vehicles, the internet of things, and fog computing. *The Eighth ACM International Workshop on Vehicular Inter-Networking (VANET)*. Association for Computing Machinery, Las Vegas
- Shi W, Cao J, Zhang Q, Li Y, Xu L (2016) Edge computing: vision and challenges. *Internet Things J IEEE* 3(5):637–646
- Edge computing industry alliance (ECC), Industrial Internet Industry Alliance (AII) (2020) Edge computing reference architecture [OL]. <http://www.eccconsortium.org/Lists/show/id/334.html>. Accessed 29 Sept 2020
- Yu Z, Jie Y, Miao L, Jinlong S, Guan G (2020) Intelligent edge computing technology based on Federated learning for video surveillance. *Acta Telecom Sin* 41(10):109–115
- Xiaoqian J, Gang C, Baibing L (2020) Application of edge computing in video surveillance. *Comput Eng Appl* 56(17):86–92
- Yi S, Hao Z, Qin Z, Li Q (2016) Fog Computing: Platform and Applications. 2015 Third IEEE Workshop on Hot Topics in Web Systems and Technologies (HotWeb). IEEE Inc., Washington D.C.
- Hu H, Shan H, Wang C, Sun T, Zhen X, Yang K, et al. (2020) Video surveillance on mobile edge networks—a reinforcement-learning-based approach. *IEEE Internet Things J* 7(6):4746–4760
- Klser A, Marszałek M, Schmid C (2010) A Spatio-Temporal Descriptor Based on 3D-Gradients. *British Machine Vision Conference*, Leeds
- Wang H, Klser A, Schmid C, Liu CL (2013) Dense trajectories and motion boundary descriptors for action recognition. *Int J Comput Vis* 103(1):60–79
- Ha J, Park J, Kim H, Park H, Paik J (2018) Violence detection for video surveillance system using irregular motion information. 2018 International Conference on Electronics, Information, and Communication (ICEIC). IEEE Inc., Honolulu
- Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. *Comput Linguist* 1(4):568
- Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional Two-Stream Network Fusion for Video Action Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, Las Vegas
- Ji S, Xu W, Yang M, Yu K (2013) 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Trans Pattern Anal Mach Intell* 35(1):221–231
- Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning Spatiotemporal Features with 3D Convolutional Networks. *IEEE International Conference on Computer Vision (ICCV)*. IEEE Inc., Santiago
- Saveliev A, Uzdiaev M, Dmitrii M (2019) Aggressive Action Recognition Using 3D CNN Architectures, 2019 12th International Conference on Developments in eSystems Engineering (DeSE). IEEE Inc., Kazan
- Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning Spatiotemporal Features with 3D Convolutional Networks, 2015 IEEE International Conference on Computer Vision (ICCV), Santiago
- Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos “in the wild”, 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE Computer Society, Miami
- Graves A, Mohamed AR, Hinton G (2013) Speech recognition with deep recurrent neural networks. *Acoustics Speech & Signal Processing. icassp.international Conference on*. IEEE Inc., Vancouver
- Hochreiter S, Schmidhuber J (1997) Long Short-Term Memory. *Neural Comput* 9(8):1735–1780
- Cho K, Van Merriënboer B, Bahdanau D, Bengio Y (2014) On the properties of neural machine translation: encoder-decoder approaches. *CoRR abs/1409.1259*:103–111
- Donahue J, et al. (2016) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Trans Pattern Anal Mach Intell* 39(4):677–691
- Jiang Y, Wu Z, Tang J, Li Z, Xue X, Chang S (2018) Modeling Multimodal Clues in a Hybrid Deep Learning Framework for Video Classification. *IEEE Trans Multimed* 20(11):3137–3147
- Wei S, Song Y, Zhang Y (2017) Human skeleton tree recurrent neural network with joint relative motion feature for skeleton based action recognition, 2017 IEEE International Conference on Image Processing (ICIP). IEEE Computer Society, Beijing
- Song S, Lan C, Xing J, Zeng W, Liu J (2018) Spatio-temporal attention-based lstm networks for 3d action recognition and detection. *IEEE Trans Image Process* 27(7):3459–3471
- Kim TS, Reiter A (2017) Interpretable 3D Human Action Analysis with Temporal Convolutional Networks, 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Honolulu. <https://doi.org/10.1109/CVPRW.2017.207>
- Fragkiadaki K, Levine S, Felsen P, Malik J (2015) Recurrent Network Models for Human Dynamics. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago. <https://doi.org/10.1109/ICCV.2015.494>
- Jain A, Zamir AR, Savarese S, Saxena A (2016) Structural-RNN: Deep Learning on Spatio-Temporal Graphs. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas. <https://doi.org/10.1109/CVPR.2016.573>
- Martinez J, Black MJ, Romero J (2017) On Human Motion Prediction Using Recurrent Neural Networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu. <https://doi.org/10.1109/CVPR.2017.497>
- Ke Q, Bennamoun M, Rahmani H, An S, Sohel F, Boussaid F (2020) Learning Latent Global Network for Skeleton-Based Action Prediction. *IEEE Trans Image Process* 29:959–970
- Kong Y, Tao Z, Fu Y (2020) Adversarial Action Prediction Networks. *IEEE Trans Pattern Anal Mach Intell* 42(3):539–553
- Lea C, Flynn MD, Vidal R, Reiter A, Hager GD (2017) Temporal Convolutional Networks for Action Segmentation and Detection. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu. <https://doi.org/10.1109/CVPR.2017.113>
- Kim TS, Reiter A (2017) Interpretable 3D Human Action Analysis with Temporal Convolutional Networks. 2017 IEEE Conference on Computer



Vision and Pattern Recognition Workshops (CVPRW), Honolulu. <https://doi.org/10.1109/CVPRW.2017.207>

40. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas. <https://doi.org/10.1109/CVPR.2016.90>
41. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You Only Look Once: Unified, Real-Time Object Detection. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas. <https://doi.org/10.1109/CVPR.2016.91>
42. Redmon J, Farhadi A (2018) YOLOv3: An incremental improvement. arXiv:1804.02767. <https://arxiv.org/abs/1804.02767>. Accessed 26 Sept 2020
43. Li W, Zhang Z, Liu Z (2020) Action recognition based on a bag of 3D points. 2020 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco. <https://doi.org/10.1109/CVPRW.2020.5543273>

### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Submit your manuscript to a SpringerOpen<sup>®</sup> journal and benefit from:**

- Convenient online submission
- Rigorous peer review
- Open access: articles freely available online
- High visibility within the field
- Retaining the copyright to your article

---

Submit your next manuscript at ► [springeropen.com](https://www.springeropen.com)