Service Action Recognition in Power Supply Business Hall with 3D-Fused ConvNet

Tongyao LIN[1], *Li OUYANG*[2], *He WEN*[1], *Dezhi XIONG*[2], *Janusz SMULKO*[3]

[1]College of Electrical and Information Engineering, Hunan University,
Lushan S Road 2, 410007 Changsha, China
[2]Hunan Province Key Laboratory of Intelligent Electrical Measurement and Application Technology,
State Grid Hunan Electric Power Company Power Supply Service Center (Metrology Center),
Lushan North Road 388, 410007 Changsha, China
[3] Department of Metrology and Optoelectronics, Faculty of Electronics, Telecommunications and Informatics,
Gdansk University of Technology, Gabriela Narutowicza 11/12, 80-233 Gdansk, Poland

linty303@163.com, 424379318@qq.com, he_wen82@126.com, dzhxiong@outlook.com, jsmulko@eti.pg.edu.pl

DOI: 10.15598/aeee.v19i1.3950

Article history: Received Sep 22, 2020; Revised Jan 04, 2021; Accepted Feb 10, 2021; Published Mar 31, 2021.

Abstract. *For the purpose of improving the service quality, video surveillance systems are widely used to standardize the service process in power supply business halls. If the employers check surveillance video to ensure predefined process of staff behaviours, it will be characterized as time-consuming. In recent years, great progress has been made in intelligent action recognition using Convolution Neural Networks (CNNs). However, due to the small range of staffs' motion and similar scene information of power supply business halls, the performance of using traditional CNNs to recognize service actions, e.g. bowing, standing and sitting, is general. For improving the recognition rate, this paper proposes a 3D-fused Convolutional Network (ConvNet) for service actions recognition, which focuses on detecting the actions in the typical scene of one staff person and one customer with a well-segmented video clip. The well-segmented video clips are sent as input to the 3D-fused ConvNet for action recognition. The 3D-fused ConvNet consists of two base learners, optical flow base learner and RGB base learner. Both learners use the Convolutional 3D (C3D) architecture. Specifically, the RGB learner can be used to capture the features of small staffs' motion while the optical flow base learner can be viewed as the key part to eliminate the influence of the background, especially in a similar scene. Furthermore, prediction scores of two base learners can be weighted by the softmax function according to the performance of each base learner. Finally, the prediction scores of the two base learners are fused to obtain the prediction result, namely the specific actions of the staffs in the videos. The experiment result shows that the proposed method achieves* 92.41 *% accuracy on the service action dataset of the power supply business hall.*

Keywords

3D convolution, action recognition, ensemble, power supply business hall.

1. Introduction

In a society with fierce market competition, quality service is the lifeblood of an enterprise. To improve the service quality in the power supply business hall, many electric companies intend to implement video recognition techniques to detect the service actions of the staffs. However, different from usual action recognition [10] and [19], service video recognition is more complicated. Most service video recognition scenes are static and similar, which means that the class of action cannot be distinguished from the background information. Moreover, the range of staffs' motion is small, and some service actions are similar. Therefore, it's hard to recognize service actions.

Most of the traditional action recognition methods are based on hand-crafted representations, such as

Motion Boundary Histograms (MBH) [3], STIPs [11] and iDT [23]. These methods are based on the features of the action to design effective algorithms to extract the motion features in the video. However, due to the lack of flexibility and scalability of hand-made representations when facing large data sets, these methods have low action recognition accuracy. In recent years, deep learning methods have been widely used to solve complex problems in the computer vision field [13], such as face recognition [9] and [25], animal recognition [22] and [26]. In action recognition, the use of learned representations with a deep learning approach has better performance than hand-made representations [18] and [24].

As one of the representative algorithms of deep learning, 2D Convolution Neural Networks (2D-CNNs) are widely used in learned feature extraction. Region convolution neural network [7] was designed for object detection. Donahue et al. [4] proposed a recurrent 2D convolutional architecture for visual understanding tasks. However, 2D-CNNs extract features from the spatial dimensions only and fail to capture temporal features which are the key to action recognition. To obtain temporal features, Ji et al. [8] developed 3D convolution, which extends the 2D convolution in the temporal dimension. Then Tran et al. [21] proposed a 3D Convolution network (C3D) using 3D convolution operations to extract features map in both spatial and temporal dimensions. Even for small movements, C3D has a reasonable recognition rate.

However, the performance of traditional CNNs with RGB frame inputs only is more dependent on the object or scene information of the training set [17]. In the power supply business hall, the background of the service video is static, which contains very little information of the scene. Therefore, relying only on CNNs with RGB inputs for service action recognition is very difficult. There are also many action recognition methods for static frames, such as SIFT [14], HOG [2]. But they cannot solve problems such as objects being occluded or are unable to extract features accurately.

To better recognize the action, as tested by the experiment [16], the optical flow is also useful in action recognition. When the observer is not moving, using optical flow can eliminate the influence of the background and other irrelevant factors. However, the accuracy of optical flow will decrease in the presence of noise [15]. Therefore, many two-stream CNN architectures that take optical flow and RGB frames as two networks' inputs were proposed to extract Spatio-temporal features. However, the two-stream network uses only a single frame of RGB images in a spatial dimension and uses stacked optical flow frames in a temporal dimension. It makes the network access to Spatio-temporal features particularly limited [6].

In this paper, the 3D-Fused ConvNet based on RGB base learner and optical flow base learner is proposed to achieve service action recognition in the power supply business hall, which is characterized by its high recognition rate. Both base learners are C3D architecture. The RGB learner can be used to capture the features of small staffs' motion while the influence of background and other irrelevant factors can be eliminated by the optical flow learner. Firstly, in order to establish a dataset, the service action videos of staffs in the power supply business hall are divided into clips according to the class of staffs' actions. Secondly, the clips are sent into the 3D-Fused ConvNet, and the RGB frames are extracted from the clips at intervals of eight frames while the Gunnar Farneback method [5] is used to obtain the optical flow frames. Thirdly, RGB frames and optical flow frames are sent into RGB learner and optical flow learner, respectively, to get the prediction scores for each action. In the next step, prediction scores of two base learners can be weighted by the softmax function according to the performance of each base learner. Then, the prediction scores of the two base learners are fused to obtain the prediction result, which is the specific actions of the staffs in the videos. Finally, some experiments are conducted to investigate the performance of RGB learners and optical flow learners. It can be concluded that the proposed method features by its high recognition rate from the experiment result.

2. Network Architecture and 3D-Fused ConvNet Method

In this section, the problem description, the optical flow method and the difference between 2D and 3D convolution will be briefly reviewed. Then the C3D architecture and the 3D-Fused ConvNet are given a description.

2.1. Problem Description

In this paper, we focus on the service action recognition in the typical power supply business hall scene of less than two people with a well-segmented video. Six service actions commonly used in the service process are selected to establish our dataset, including sitting, introducing service content, standing, bowing, submitting materials and shaking hands. Among these actions, the detection of submitting materials and shaking hands are the most challenging tasks since they have the same backgrounds. Specifically, the submitting materials and shaking hands have the same features in translation motion and arm pendulu-like motion, as shown in Fig. 1. In addition, the range of movement of some actions is quite small, such as standing, bowing, and introducing service content, which also leads to the recognition of service actions is challenging.

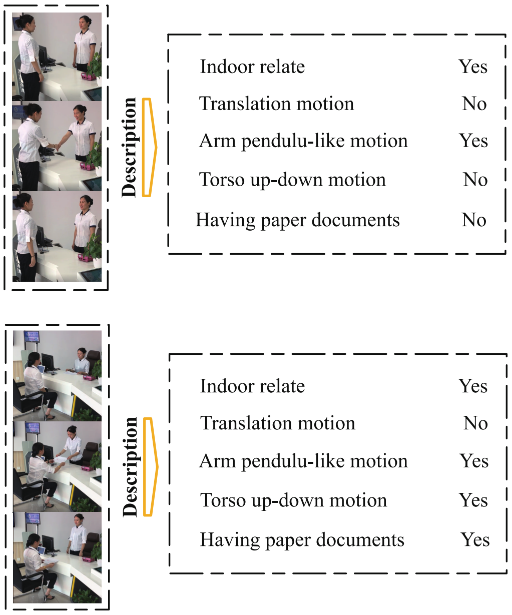

Fig. 1: The features of service actions of submitting materials and shaking hands.

2.2. Optical Flow

Optical flow is the motion vector field of pixels in a two-dimensional image, which is often used to detect and estimate the object. The optical flow method attempts to calculate the motion vector field between two image frames, which are taken at the time t and $t + \Delta t$ at each pixel's position. At position (x, y, t), assuming the intensity $A(x, y, t)$ of pixels following brightness constancy constraint can be given:

$$A(x, y, t) = A(x + \Delta x, y + \Delta y, t + \Delta t). \tag{1}$$

Perform Taylor approximation on the right-hand side of Eq. (1) and divide dt on both sides of the equation. The optical flow equation is obtained by:

$$\frac{dA}{dx}u + \frac{dA}{dy}v + \frac{dA}{dt} = 0, \tag{2}$$

where $u = \frac{dx}{dt}$ and $v = \frac{dy}{dt}$. Also, $\frac{dA}{dx}$ is the image gradient along the horizontal axis, $\frac{dA}{dy}$ is the image gradient along the vertical axis, and $\frac{dA}{dt}$ is along the time. However, there is only one equation but two unknowns. In this paper, the Gunnar Farneback method [5] is chosen as an additional constraint to calculate the motion vector field. The algorithm is briefly reviewed as following in the Alg. 1, where $\odot$ represents the Hadamard product of two vectors. Then optical flow frames can be obtained by coloring motion vector field with Munsell color system. Different colors, shades of colors of the optical flow frames indicate the different motion intensity and direction of the object, respectively. Thus, the optical flow frames indicate the information about the movement of objects while the influence of background, staffs' clothes and other irrelevant factors on action recognition are eliminated.

Algorithm 1 Dense Optical Flow (Gunnar Farneback Method).

Require: prev: the RGB frames.
Ensure: $\vec{\delta}$: a parameter vector for each pixel in a single image.

```
$\vec{u}$ ←prev //Convert to grayscale
$f(\vec{u}) \leftarrow \vec{u}^T \mathbf{A} \vec{u} + \vec{b}^T \vec{u} + c$ //Binomial modeling
$f(\vec{u})$ parameterized to get $(b_1, b_2, \ldots, b_6) \cdot \vec{e}$
$\vec{B} \leftarrow \vec{a} \odot (b_1, b_2, \ldots, b_6)$ //Add weight
$\vec{\theta} \leftarrow$ The dual conversion $\vec{B}$
$\vec{\delta} \leftarrow \vec{\theta}^{-1}$
```

2.3. 2D Convolution and 3D Convolution

2D-CNNs have strong feature extraction capabilities in operating two-dimensional images. The essence is to use each convolution kernel to calculate the feature map. The value at position (x, y) on the j-th feature map in the i-th layer can be computed as:

$$v_{ij}^{xy} = f\left(\sum_{m} \sum_{p=0}^{P_i-1} \sum_{q=0}^{Q_i-1} w_{ijm}^{pq} v_{(i-1)m}^{(x+p)(y+q)} + b_{ij}\right), \tag{3}$$

where b_{ij} is the bias of the feature map; w_{ijm}^{pq} indicates the weights of the kernel connected to the m-th feature map in the previous layer; P_i, Q_i is height and width of a 2D kernel; $f(\bullet)$ is the nonlinear activation function.

However, as shown in Fig. 2, a convolution operation only extracts the features of one frame in the 2D convolution. Hence, 2D convolution contains no time information which cannot be obtained through a single frame. To better extract temporal information, 3D-CNNs stack multiple consecutive frames into a cube and use 3D convolution kernel to obtain Spatio-temporal information. The value at position (x, y, z) on the j-th feature map in the i-th layer can be computed as shown in Eq. (4), where w_{ijm}^{pqr} indicates the weights of the kernel connected to the m-th feature map in the previous layer; P_i, Q_i are the height and width of a 3D kernel; R_i is the size of the 3D kernel along a temporal dimension. Figure 2 shows the comparison of 2D convolution and 3D convolution operations. H and W represent the height and width of the image frame, respectively, and T represents the timeline. The 2D convolution can only extract features on

$$v_{ij}^{xyz} = f\left(\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)} + b_{ij}\right). \tag{4}$$

a single RGB frame, while the 3D convolution can extract features on multiple consecutive RGB frames at the same time.

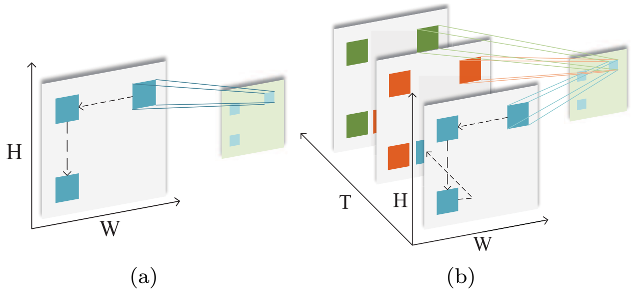

Fig. 2: Comparison of 2D convolution (a) and 3D convolution (b) operations.

2.4. C3D Architecture

The C3D is the most representative 3D-CNNs. 16 consecutive service action frames with the size of 112×112 are considered as inputs to the C3D architecture. Table 1 shows the C3D architecture. C is the number of channels, D is the number of frames. There are 8 convolutional layers, 5 pooling layers, 2 fully-connected layers and a softmax output layer.

Tab. 1: The architecture and dimension of C3D.

Name	**Kernel dims (R×P×Q)**	**Output dims (C×D×H×W)**
Conv1	$3 \times 3 \times 3$	$64 \times 16 \times 112 \times 112$
Max-pooling1	$1 \times 2 \times 2$	$64 \times 16 \times 56 \times 56$
Conv2	$3 \times 3 \times 3$	$128 \times 16 \times 56 \times 56$
Max-pooling2	$2 \times 2 \times 2$	$128 \times 8 \times 28 \times 28$
Conv3a	$3 \times 3 \times 3$	$256 \times 8 \times 28 \times 28$
Conv3b	$3 \times 3 \times 3$	$256 \times 8 \times 28 \times 28$
Max-pooling3	$2 \times 2 \times 2$	$256 \times 4 \times 14 \times 14$
Conv4a	$3 \times 3 \times 3$	$512 \times 4 \times 14 \times 14$
Conv4b	$3 \times 3 \times 3$	$512 \times 4 \times 14 \times 14$
Max-pooling4	$2 \times 2 \times 2$	$512 \times 2 \times 7 \times 7$
Conv5a	$3 \times 3 \times 3$	$512 \times 2 \times 7 \times 7$
Conv5b	$3 \times 3 \times 3$	$512 \times 2 \times 7 \times 7$
Max-pooling5	$2 \times 2 \times 2$	$512 \times 1 \times 4 \times 4$
1×1 conv	-	8192
Fc6	-	4096
Fc7	-	4096

In the convolutional layers, all kernels of 3D convolution filters are $3 \times 3 \times 3$ with stride $1 \times 1 \times 1$. The nonlinear activation function is Rectified Linear Unit (ReLU). It can effectively alleviate the problem of vanishing gradient. Here is ReLU definition:

$$\mathrm{ReLU} = \begin{cases} 0 & x < 0 \\ \max(0, x) & \text{else} \end{cases}. \tag{5}$$

Next, the 3D pooling layer is followed. It is an extension of 2D pooling in the time dimension. It stacks feature maps across time and applies max-pooling to shrink this spatiotemporal cube with a 3D pooling cube. The pooling filters of the size: $1 \times 2 \times 2$ and $2 \times 2 \times 2$, which can preserve time information. The pooling layer can be defined as:

$$\alpha_j^i = \beta_j^i \max\left(\alpha_j^{(i-1)}\right) + b_j^i, \tag{6}$$

where $\max(\bullet)$ computes the max values over a neighbourhood in each feature map. β_i^j and b_i^j represent multiplicative bias and an additive bias.

The last two fully-connected layers contain 4096 output units which will be sent to the softmax classifier for classification. The value of softmax can be defined as:

$$Y_{gn} = \frac{\exp\left(x_{gt}\right)}{\sum_{t=1}^{N} \exp\left(x_{gt}\right)}, \tag{7}$$

where Y_{gn} represents the probability that the g-th sample belongs to the n-th class of action. x_{gt} represents the t-th element in the output vector of the g-th sample after passing through the classification layer.

To train the C3D model, the stochastic gradient descent [1] is used to minimize the difference between true value and the actual output of the network. The loss function is defined as follows:

$$L = -\frac{1}{G}\sum_{i=1}^{G}\sum_{j=1}^{N} \ln\left(Y_{gn}\overline{Y}_{gn}\right), \tag{8}$$

where G is the total number of samples, $\overline{Y}_{gn}$ represents the true probability that the g-th sample belongs to the n-th class of action.

2.5. 3D-Fused ConvNet

Figure 3 shows an overview of the whole 3D-Fused ConvNet architecture. The architecture consists of three parts: an input layer, base learner part and class score fusion.

In the input layer, the optical flow frames are extracted from the video, while the RGB frames are obtained at intervals of eight frames. Then the base

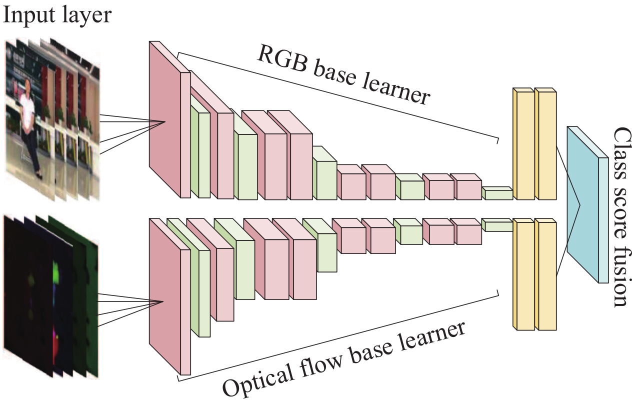

Fig. 3: 3D-Fused ConvNet architecture. The pink boxes, the green boxes, the yellow boxes and the blue box represent the convolutional layer, pooling layer, softmax layer and class score fusion layer.

learner part is followed, the input of the first base learner is RGB frames, and the input of the second base learner is the optical flow. Each base learner is trained for better extract features. In class score fusion, according to the recognition accuracy rate of each base learner for each class of service action, the weight $\vec{\lambda}$ is determined by the softmax formula. To increase the weight of the base learner with high accuracy, the recognition rate of each action is multiplied by 10, and then the softmax formula is used:

$$\vec{\lambda} = \left(\frac{\exp(z_{11})}{\sum_{t=1}^{2} \exp(z_{1t})}, \frac{\exp(z_{21})}{\sum_{t=1}^{2} \exp(z_{2t})}, \cdots, \frac{\exp(z_{n1})}{\sum_{t=1}^{2} \exp(z_{nt})} \right), \tag{9}$$

where z_{nt} represents 10 times the recognition accuracy rate of the t-th base learner for the n-th action recognition. The prediction score of each base learner will be multiplied by weight λ to obtain the final prediction score. The 3D-Fused convolution network uses the service action class with the highest prediction score as the action recognition result. The final prediction score can be computed as:

$$\vec{Z}_f = \vec{\lambda} \odot \vec{z}_1 + \left(\vec{E} - \vec{\lambda} \right) \odot \vec{z}_2, \tag{10}$$

where $\odot$ represents the Hadamard product of two vectors; $\vec{z}_1 = (z_{11}, z_{21}, \ldots, z_{n1})$, $\vec{z}_2 = (z_{12}, z_{22}, \ldots, z_{n2})$. $\vec{E}$ is vector with all elements equal to 1 and with the same dimensions as $\vec{z}_1$ and $\vec{z}_2$. Also, $\vec{z}_1$ and $\vec{z}_2$ represent the prediction score of two base learners, respectively; n represents the number of class of service action recognition; $\vec{Z}_f$ represents the score of final prediction.

3. Experiments

In this section, we first introduce our server action dataset provided by Hunan Electric Power Metering Center, and then the training details are presented. In the last, the performance of two base learners and the 3D-Fused ConvNet are evaluated.

3.1. Dataset

The service action videos of staffs are divided into clips with video editing software according to the staffs' movements to establish our dataset. The method of C3D starts by capturing the appearance of a single object for the first few frames of a video, then focusing on the detection of actions of this object [21], which leads to the inability to recognize the actions of multiple people simultaneously. If the videos contain multiple gestures or multiple people, we use video software to cut the video into a scene containing only one gesture or no more than two people.

As shown in Tab. 2, the dataset contains 998 clips, of which 896 clips are recorded in the Shaoshan power supply business hall, and the rest are recorded in different scenes. Like UCF101 [19] dataset, we convert all clip sizes to 320 × 240 pixels spatial resolution.

Tab. 2: Summary of characteristics of our dataset.

Attribute	Value
Actions	6
Clips	998
Groups per Action	5–7
Clips per Group	20–30
Mean Clips length	3.2 s
Frame Rate	30 fps
Resolution	320 × 240

There are six common service actions in total. Figure 4 shows the class of server actions, such as bowing, shaking hands, introducing service content to users, sitting down, standing and submitting materials. Each action class that has at least 100 video examples is recorded. In order to evaluate our algorithm, we randomly divide the 998 clips into training set and testing set according to the ratio of 4 : 1.

3.2. Implementation Details

Firstly, while calculating dense optical flow with an i7-7700HQ CPU and OpenCV library to obtain the optical flow frames at 16.8 frames per second (fps), the service action clips are read at 8-frame interval to obtain RGB frames. Secondly, after the optical flow frames are obtained, the RGB frames and the optical flow frames are input to 3D-fused ConvNet at the same time. The 3D-fused ConvNet is trained and tested with a GPU (2080Ti) under the implementation of TensorFlow and Linux. Specifically, in order to avoid overfitting, the batch size is set to 10, and network weights are initialized with Sports-1M pre-trained models. The

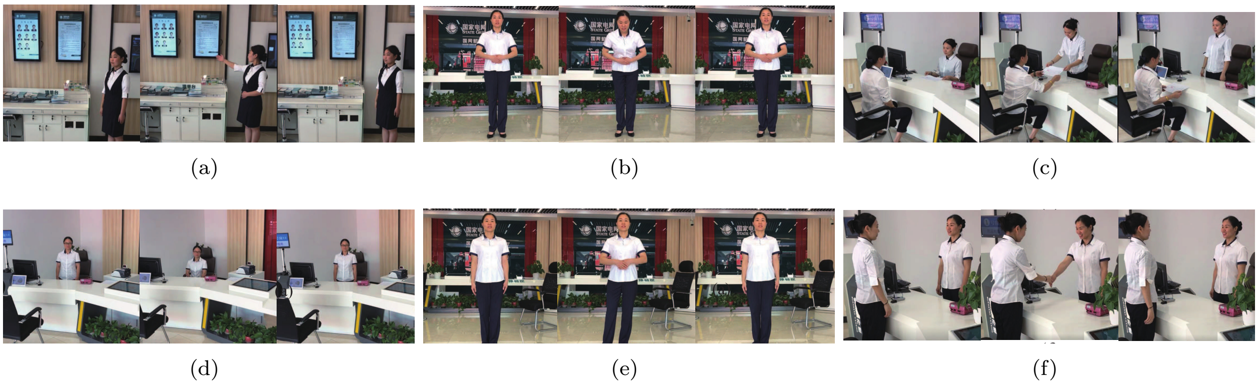
(a) (b) (c) (d) (e) (f)

Fig. 4: Six actions included in our dataset shown with three sample frames. (a), (b), (c), (d), (e) and (f) are introducing, bowing, submitting materials, sitting down, standing and handshaking, respectively.

initial learning rate is 0.0001, and the moving average decay is 0.999. The Adam [12] optimization algorithm is used to update the network parameters. The cross-entropy loss is employed to backpropagate gradients. Finally, the clips of testing set are sent into the 3D-fused ConvNet to get the prediction result. The average speed of calculating the final result is 12.4 fps.

Tab. 3: Comparison of action recognition accuracy in different maximum step size.

Maximum step sizes	Accuracy of RGB learner [21]	Accuracy of optical flow learner [20]
500	83.48 %	38.83 %
100	88.39 %	81.46 %
1500	90.65 %	79.68 %
3000	87.94 %	77.89 %
5000	88.83 %	78.13 %
8000	88.39 %	81.70 %

3.3. Experiment and Results

In this part, the selection of maximum step sizes of each base learner is analyzed firstly. Then the analysis for the performance in changing scene and confusion matrix on our dataset will be shown. Finally, the accuracy of the 3D-Fused ConvNet will be compared with other methods.

To evaluate the performance of the two base learners under different maximum step sizes, the experiment sets the maximum step size of 500, 1000, 1500, 3000, 5000 and 8000. Table 3 shows the recognition rate of two base learners on our dataset.

The accuracy rate does not increase with the increase of the number of step sizes. RGB learner and optical flow learner achieve the best performance when the maximum number of step sizes are 1500 and 8000, respectively. By considering the correct rate and the time spent, 1500 and 1000 are finally selected as the maximum step size of the two base learners in all the next experiments. Table 3 also indicates that the performance of RGB learner is significantly better than optical flow learner; this is because RGB frames contain more information than optical flow. Although the result of RGB learner is much better than optical flow learner, RGB learner cannot be used only for service action recognition. This conclusion will be proved in the subsequent two experiments.

To further illustrate the effectiveness of two base learners, the performance of each class of action is demonstrated, and the impact of changes scene will be shown in the next. Figure 5 and Fig. 6 show the confusion matrices of RGB learner and optical flow learner on our dataset. The number in the box indicates the probability that the predicted label is the corresponding true label. The performance of RGB learner is well in recognizing the standing and the introducing states, even though the movement of the standing is minimal. However, the RGB learner is weaker than the optical flow learner in terms of sitting down and introducing action recognition. The Optical flow learner, which has 96 % and 94 % in recognizing the sitting down and the introducing state, respectively, is better than the RGB learner. It is because dense optical flow, which computes the vector for every pixel of each frame, is more sensitive to movements with large amplitudes. However, the optical flow will confuse movements with small or similar actions. For example, the optical flow has 76 % recognition rate in submitting, and there is an 11 % probability that the submission will be recognized as a handshake.

To explore the impact of changes scene on the two base learners, the 896 clips recorded in the electrical hall are put into training set, and 108 clips recorded in different scenes were put into the testing set. As Fig. 7 shows, the accuracy of the RGB learner and the optical flow learner is 47.22 % and 60.18 %, respectively. When changing to an untrained scene, the accuracy of the RGB learner descends sharply, while the accu-

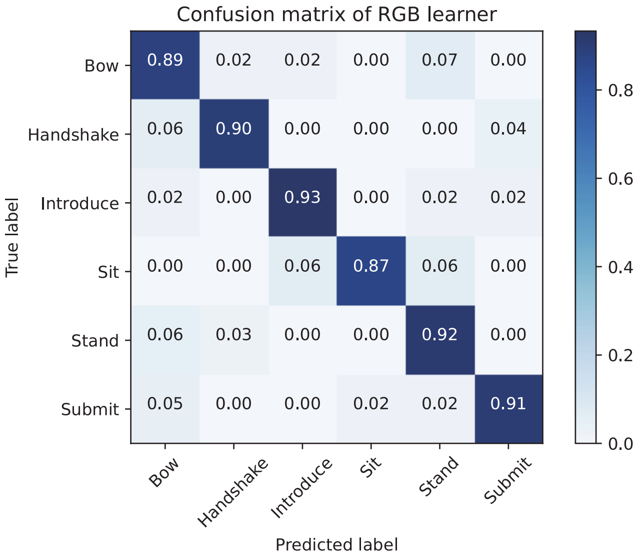

Fig. 5: Confusion matrix of RGB learner.

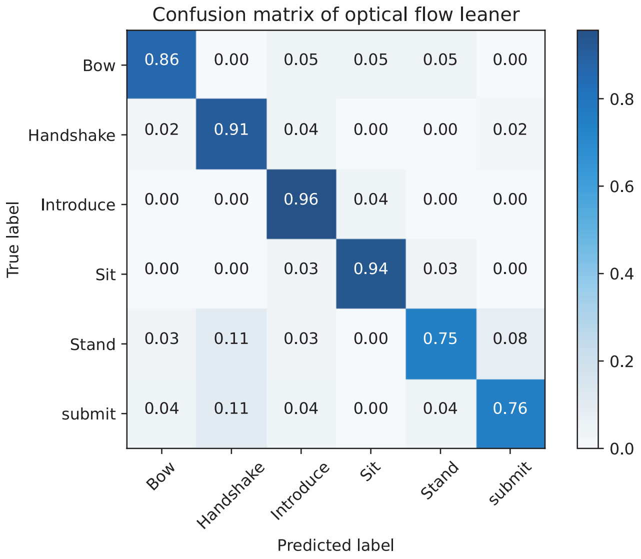

Fig. 6: Confusion matrix of optical flow learner.

racy of the optical flow learner is relatively stable. The RGB base learner is weak in generalization ability and relies too much on the information in the training set. Our method, the 3D-Fused ConvNet, which combines two base learners, has better stability in the face of untrained scenes.

Based on the above two experiments, both base learners have their advantages and disadvantages. The RGB learner has an excellent performance on service action recognition and even has a very high recognition rate for small-scale actions like standing. The optical flow learner has the advantages of high recognition rate in large-scale works and good stability. Our method adjusts the weights of the two base learners to give play to the advantages of the base learners and reduces their shortcomings. Figure 8 shows the comparison between our method and other methods in each class of action recognition accuracy rate. Unlike other ways that have low recognition rates in some actions, our approach has high accuracy and a stable recognition rate for each service action.

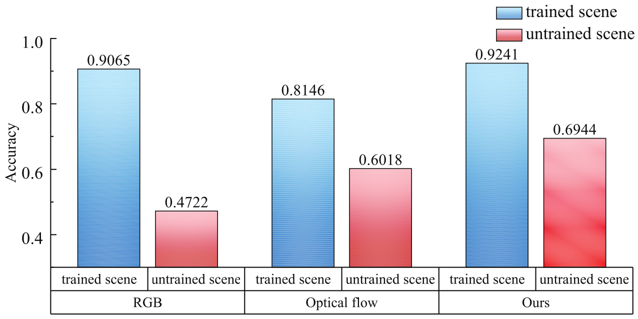

Fig. 7: Comparison of each learner with the trained scene and each learner with the untrained scene.

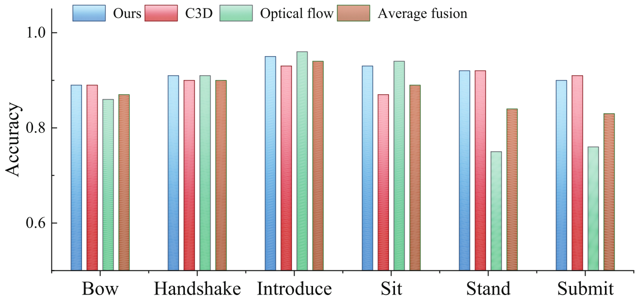

Fig. 8: Comparison between our method and other methods in each class of action recognition accuracy rate.

Above all, after the exploration and analysis of the 3D-Fused ConvNet, our result is shown in Tab. 4. The recognition accuracy of the 3D-Fused ConvNet is better than other methods. The experimental results also prove the effectiveness of the 3D-Fused ConvNet.

Tab. 4: Comparison of action recognition accuracy with other methods.

Method	Accuracy
C3D [21]	90.65 %
Optical flow [20]	81.70 %
Average fusion [18]	87.05 %
Ours	92.41 %

4. Conclusion

This paper establishes a dataset with six service actions that are commonly used in the service process and proposes a 3D-Fused ConvNet that effectively combines RGB information and optical flow information for the accurate recognition of service actions in a typical power supply business hall scene of less than two people. The experiments show that the RGB learner can capture the small range of service actions, and the optical flow learner can eliminate the influence of background and other irrelevant factors. In future works, we will explore the specific performance of the 3D-fused ConvNet in a scene that contains multiple people or in a different room.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China under Grant 61771190 and in part by the Natural Science Foundation of Hunan Province under Grant 2019JJ20001.

Author Contributions

T.L. and L.O. set up the dataset. T.L. and He Wen were involved in planning and supervised the work. T.L. and L.O. processed the dataset, performed the analysis, drafted the manuscript and designed the figures. T.L. and L.O. measurements the dataset with 3D-fused CNN. T.L. and He Wen analyzed the experimental result. D. X. and J.S. aided in interpreting the results and worked on the manuscript. All authors discussed the results and commented on the manuscript.

References

[1] BOTTOU, L. Stochastic gradient descent tricks. *Neural Networks: Tricks of the trade*. 2nd ed. Berlin: Springer, 2012, pp. 421–436. ISBN 978-3-642-35289-8.

[2] DALAL, N. and B. TRIGGS. Histograms of oriented gradients for human detection. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*. San Diego: IEEE, 2005, pp. 886–893. ISBN 0-7695-2372-2. DOI: 10.1109/CVPR.2005.177.

[3] DALAL, N., B. TRIGGS and C. SCHMID. Human Detection Using Oriented Histograms of Flow and Appearance. In: *European Conference on Computer Vision ECCV 2006: Computer Vision – ECCV 2006*. Graz: Springer, 2006, pp. 428–441. ISBN 978-3-540-33835-2. DOI: 10.1007/11744047_33.

[4] DONAHUE, J., L. A. HENDRICKS, M. ROHRBACH, S. VENUGOPALAN, S. GUADARRAMA, K. SAENKO and T. DARRELL. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2017, vol. 39, iss. 4, pp. 677–691. ISSN 1939-3539. DOI: 10.1109/TPAMI.2016.2599174.

[5] FARNEBACK, G. Two-Frame Motion Estimation Based on Polynomial Expansion. In: *Scandinavian Conference on Image Analysis SCIA 2003: Image Analysis*. Halmstad: Springer, 2003, pp. 363–370. ISBN 978-3-540-45103-7. DOI: 10.1007/3-540-45103-X_50.

[6] FEICHTENHOFER, C., A. PINZ and A. ZISSERMAN. Convolutional Two-Stream Network Fusion for Video Action Recognition. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas: IEEE, 2016, pp. 1933–1941. ISBN 978-1-4673-8851-1. DOI: 10.1109/CVPR.2016.213.

[7] GIRSHICK, R. B., J. DONAHUE, T. DARRELL and J. MALIK. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. Columbus: IEEE, 2014, pp. 580–587. ISBN 978-1-4799-5118-5. DOI: 10.1109/CVPR.2014.81.

[8] JI, S., W. XU, M. YANG and K. YU. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2013, vol. 35, iss. 1, pp. 221–231. ISSN 1939-3539. DOI: 10.1109/TPAMI.2012.59.

[9] KAMENCAY, P., M. BENCO, T. MIZDOS and R. RADIL. A New Method for Face Recognition Using Convolutional Neural Network. *Advances in Electrical and Electronic Engineering*. 2017, vol. 15, iss. 4, pp. 663–672. ISSN 1804-3119. DOI: 10.15598/aeee.v15i4.2389.

[10] KUEHNE, H., H. JHUANG, E. GARROTE, T. POGGIO and T. SERRE. HMDB: A large video database for human motion recognition. In: *2011 International Conference on Computer Vision*. Barcelona: IEEE, 2011, pp. 2556–2563. ISBN 978-1-4577-1102-2. DOI: 10.1109/ICCV.2011.6126543.

[11] LAPTEV, I. On Space-Time Interest Points. *International Journal of Computer Vision*. 2005, vol. 64, iss. 2, pp. 107–123. ISSN 1573-1405. DOI: 10.1007/s11263-005-1838-7.

[12] LIU, L. and L. SHAO. Learning discriminative representations from RGB-D video data. In: *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence (IJCAI)*. Beijing: AAAI Press, 2013, pp. 1493–1500. ISBN 978-1-57735-633-2. DOI: 10.5555/2540128.2540343.

[13] LIU, G., B. HE, J. WU and Z. LIN. Deep Learning Image Classification Network for Visual Inspection and Its Application in Components Quality Test. *China Measurement & Testing Technology.* 2019, vol. 45, iss. 7, pp. 1–10. ISSN 1672-4984.

[14] LOWE, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision.* 2004, vol. 60, iss. 2, pp. 91–110. ISSN 1573-1405. DOI: 10.1023/B:VISI.0000029664.99615.94.

[15] OREIFEJ, O. and Z. LIU. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In: *2013 IEEE Conference on Computer Vision and Pattern Recognition.* Portland: IEEE, 2013, pp. 716–723. ISBN 978-1-5386-5672-3. DOI: 10.1109/CVPR.2013.98.

[16] SEVILLA-LARA, L., Y. LIAO, F. GUNEY, V. JAMPANI, A. GEIGER and M. J. BLACK. On the Integration of Optical Flow and Action Recognition. In: *German Conference on Pattern Recognition (GCPR).* Stuttgart: Springer, 2018, pp. 281–297. ISBN 978-3-030-12939-2. DOI: 10.1007/978-3-030-12939-2_20.

[17] SHAO, D., Y. ZHAO, B. DAI and D. LIN. FineGym: A Hierarchical Video Dataset for Fine-Grained Action Understanding. In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* Seattle: IEEE, 2020, pp. 2616–2625. ISBN 978-1-7281-7168-5. DOI: 10.1109/CVPR42600.2020.00269.

[18] SIMONYAN, K. and A. ZISSERMAN. Two-stream convolutional networks for action recognition in videos. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1 (NIPS).* Montreal: MIT Press, 2014, pp. 568–576. DOI: 10.5555/2968826.2968890.

[19] SOOMRO, K., A. R. ZAMIR and M. SHAH. *UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild.* Orlando: Center for Research in Computer Vision, University of Central Florida, 2012. Available at: https://arxiv.org/pdf/1212.0402.pdf.

[20] TAKAMINE, A., Y. IWASHITA and R. KURAZUME. First-person activity recognition with C3D features from optical flow images. In: *2015 IEEE/SICE International Symposium on System Integration (SII).* Nagoya: IEEE, 2015, pp. 619–622. ISBN 978-1-4673-7242-8. DOI: 10.1109/SII.2015.7405050.

[21] TRAN, D., L. BOURDEV, R. FERGUS, L. TORRESANI and M. PALURI. Learning Spatiotemporal Features with 3D Convolutional Networks. In: *2015 IEEE International Conference on Computer Vision (ICCV).* Santiago: IEEE, 2015, pp. 4489–4497. ISBN 978-1-4673-8391-2. DOI: 10.1109/ICCV.2015.510.

[22] TRNOVSZKY, T., P. KAMENCAY, R. ORJESEK, M. BENCO and P. SYKORA. Animal Recognition System Based on Convolutional Neural Network. *Advances in Electrical and Electronic Engineering.* 2017, vol. 15, iss. 3, pp. 517–525. ISSN 1804-3119. DOI: 10.15598/aeee.v15i3.2202.

[23] WANG, H. and C. SCHMID. Action Recognition with Improved Trajectories. In: *2013 IEEE International Conference on Computer Vision.* Sydney: IEEE, 2013, pp. 3551–3558. ISBN 978-1-4799-2840-8. DOI: 10.1109/ICCV.2013.441.

[24] WANG, L., Y. QIAO AND X. TANG. Action recognition with trajectory-pooled deep-convolutional descriptors. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* Boston: IEEE, 2015, pp. 4305–4314. ISBN 978-1-4673-6964-0. DOI: 10.1109/CVPR.2015.7299059.

[25] ZHANG, J., L. TANG, A. MINGOTTI, L. PERETTO and H. WEN. Analysis of White Noise on Power Frequency Estimation by DFT-Based Frequency Shifting and Filtering Algorithm. *IEEE Transactions on Instrumentation and Measurement.* 2020, vol. 69, iss. 7, pp. 4125–4133. ISSN 1557-9662. DOI: 10.1109/TIM.2019.2941290.

[26] ZHANG, J., H. WEN and L. TANG. Improved Smoothing Frequency Shifting and Filtering Algorithm for Harmonic Analysis With Systematic Error Compensation. *IEEE Transactions on Industrial Electronics.* 2019, vol. 66, iss. 12, pp. 9500–9509. ISSN 1557-9948. DOI: 10.1109/TIE.2019.2892664.

About Authors

Tongyao LIN was born in Fujian, China. He is currently pursuing the M.Sc. degree in College of Electrical and information engineering, Hunan University. His research interests include action recognition and deep learning.

Li OUYANG was born in Shaoyang, China. She received the master's degree in electrical engineering from Hunan University in 2018. Her research interests include digital power supply service and computer vision.

He WEN (corresponding author) was born in Hunan, China, in 1982. He is currently a Full Professor with the College of Electrical and Information Engineering, Hunan University, where he is also the Deputy Director of the Hunan Province Key Laboratory of Intelligent Electrical Measurement and Application Technology. His present research interests include electrical contact reliability, wireless communications, power system harmonic measurement and analysis, power quality, and digital signal processing. Dr. Wen is an Associate Editor of the IEEE Transactions on Instrumentation and Measurement, and a member of the Editorial Board of Fluctuation and Noise Letters.

Dezhi XION was born in Hubei, China. He is currently a high engineer with State Grid Hunan Electric Power Company Power Supply Service Center (Metrology Center), Changsha, China. His research interests include power supply and electricity metering.

Janusz SMULKO was born in Poland. He is currently a professor with the Faculty of Electronics, Telecommunications and Informatics, Gdansk University of Technology, Gdansk, Poland. His research interests include sensors and signal processing.