Preprints of the 21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

A Gesture Cognition Strategy for High-speed Train Drivers on Reconstructed Multiple Views

Yueyue Mu * **Xiaoyong Zhang** * **Chenglong Wang** * **Shuo Li** **
Yingze Yang * **Weirong Liu** * **Jun Peng** *

* *School of Computer Science and Engineering, Central South University, Changsha 410083, China, (Corresponding author e-mail: zhangxy@csu.edu.cn).*
** *School of Electrical and Information Engineering, Changsha University of Science and Technology, Changsha 410114, China, (e-mail: lishuo@csust.edu.cn).*

Abstract: The driver of high-speed trains usually is required to perform certain gestures to confirm the signals before implementing some operations, which is an essential validation for driving safety. However, the accuracy of gesture recognition is difficult to guarantee due to the jamming background and limited perspective. In this paper, the features of the side view and vertical view are integrated to assist classification decisions. Firstly, point clouds of the gesture are generated with RGB-D data and then projected onto two orthogonal planes to reconstruct the side and vertical view of the gesture. Secondly, multiple-view 3D Convolution Neural Network architecture is proposed with three branches of Convolution Neural Network. Combined with the front view obtained by frame difference, the model learns convolution features from three aspects of the gesture. Further, multiple-view classification results are adaptively fused to acquire the final decision. Experiments show that our approach is superior to the state-of-the-art gesture recognition methods on challenging dataset.

Keywords: Driver Gesture Congnition, Train Operation Safety, Human-centred Computing, Modeling of Human Performance, Cognitive Systems Engineering

1. INTRODUCTION

In order to achieve the safety goal, the driver must immediately make the corresponding gesture when catching sight of the pavement sign or receiving the command from the control centre, see Wang et al. (2012). The detection of driver's gesture can be described as the process of computer simulating human cognitive system to interpret gesture, which plays a vital role in safe driving environment.

The use of gestures indeed converts interaction technology from machine-centred to human-centred, see Rautaray and Agrawal (2015). Research on gesture recognition is of profound significance, e.g., Gustin et al. (2016), Sharma et al. (2012) and Asadi-Aghbolaghi et al. (2017).

As the advent of depth sensors, like Microsoft Kinect, gesture recognition based on somatosensory equipment has become a research trend, e.g., Da Gama et al. (2016), Plouffe and Cretu (2015) and Pramod et al. (2015). However, static pictures have a vastly different interpretation in many scenarios. Therefore, existing research mainly focuses on feature extraction of dynamic gestures.

Dynamic gesture regards hand movements as an orderly sequence of the time dimension. In depth-based gesture recognition, the toughest challenge is to acquire the temporal feature. The combination of 2D Convolution Neural Network(2D CNN) and Recurrent Neural Network is usually used for dealing with the recognition task mentioned above. The convolution kernel is utilized to scan the entire image to obtain the intra-feature of each frame, and then the inter-feature is captured with the help of the memory function of the Recurrent Neu-ral Network. The drawback of this method is that it is constrained by 2D input. Compared with 2D Convolution Neural Network, 3D Convolution Neural Network has a stronger ability to capture spatiotemporal features, which can be modelled simultaneously in both two dimensions, see e.g., Tran et al. (2015). The multi-stream structure is fully utilized in the scene where multiple disjoint features are merged frequently. Dual-stream 3D Convolution Neural Network is an effective way to solve the problem of dynamic gesture recognition. Two subnetworks train image sequences of different resolutions or forms respectively and then fuse the outputs to get the classification results, e.g., Molchanov et al. (2015). The possible drawbacks of multi-stream structure are that: (i) since the same gesture shows special effects from different perspectives, it is difficult for a single view to get full-scale information of gestures, (ii) model based on a single view lacks certain robustness, and (iii) environmental interference usually affects the capture of hand features.

⋆ This work is partially supported by National Natural Science Foundation of China (Grant Nos. 61873353 and 61672539). Corresponding author: Xiaoyong Zhang, E-mail: zhangxy@csu.edu.cn.

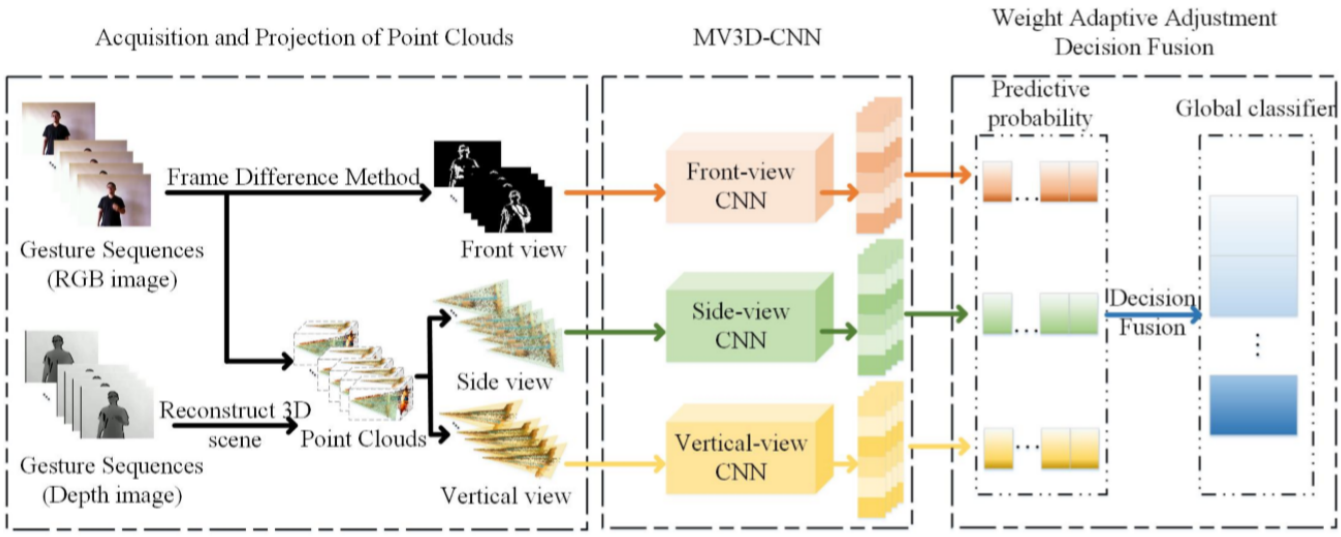

Fig. 1. The structure of the proposed method, MV3D-CNN. The inputs are RGB and depth modalities and the output is global classification result. It mainly consists of three parts: Acquisition and projection of point cloud, simple architecture of MV3D-CNN and decision fusion.

The solution to the above problems is the novel CNN architecture proposed in Fig. 1. Three reasons are given here: one is that it can make better use of depth information which can easily remove the faster environmental background and eliminate the interaction of environment and integrate multiple dimensional features to comprehend gesture significance. The second is that MV3D-CNN overcomes the problem of the lack of robustness when using a single view, and the other is to reduce the cost of computation and transform the three-dimensional problem into a two-dimensional case.

The contributions of this paper are three-fold.

• First, we propose MV3D-CNN architecture with better generalization ability. By means of implicit learning, we capture detailed and effective spatiotemporal features from the front, side and vertical views. The final classification results are obtained by combining the complementary advantages of each model with the weight adaptive adjustment decision fusion(WAADF) strategy from diff erent perspectives.

• Second, 3D point clouds are converted by video data in RGB-D format, and then multi-angle gesture projection is obtained from point cloud to represent the spatial state of gestures preferably.

• Third, our present model shows promising performance in ChaLearn LAP Isolated Gesture Dataset (IsoGD), which proves that the prospect of human-computer interaction based on gesture is considerable.

The rest of this paper is organized as follows: Section II introduces the generation of point cloud and the acquisition of gesture projection in three orthogonal planes. Section III describes the overall architecture and optimization model of MV3D-CNN for dynamic gesture recognition. In Section IV, the performance of the proposed method is evaluated with simulation results. Finally, this paper is concluded in Section V.

2. THE GENERATION OF MULTIPLE VIEWS FOR GESTURE

In this section, we project the target point cloud data onto three-dimensional orthogonal planes. Then we use the multi-perspective information contained in the front, side and vertical direction to regress the dynamic gesture. Therefore, the spatial information of gestures can be reflected vividly. Point cloud, which is obtained from RGB-D data, contains three-dimensional coordinate and color information of each sampling point.

2.1 Generation of Point Clouds

We combine two data formats of gesture video, RGB-D data, to generate point cloud data to reconstruct three-dimensional scene of the gesture. The point cloud is a sort of data that is suitable for three-dimensional scene understanding. RGB format and depth format of the original video sequence are illustrated in Fig. 2(a) and Fig. 2(b).

Given the RGB-D data, the pixel coordinates (u,v) of any point P in the RGB image, as well as z ,which represents the distance between P and sensing equipment can be obtained simultaneously. In this paper, we establish a three-dimensional sensing coordinate system with the camera as the origin.

The horizontal direction, the vertical direction, and the direction perpendicular to the camera are respectively recorded as X-Axis, Y-Axis, and Z-Axis. The point cloud data in the three-dimensional sensing coordinate system is acquired from RGB-D data. Assuming that the focus of the camera is f, we can get the coordinate (X_{pc}, Y_{pc}, Z_{pc}) of the point cloud in sensing coordinate system, which can be calculated as:

$$X_{pc} = \frac{Z_{pc}\,(u - c_x)}{f}, \tag{1}$$

$$Y_{pc} = \frac{Z_{pc}\,(v - c_y)}{f}, \tag{2}$$

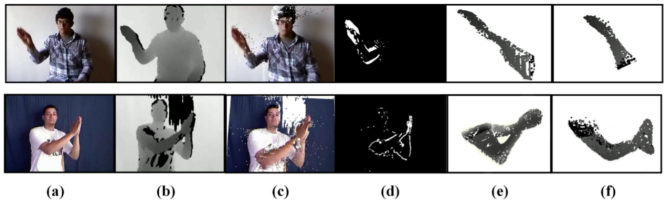

Fig. 2. The different presentation forms of two gesture.

$$Z_{pc} = d, \tag{3}$$

where (u,v) and (c_x, c_y) represent the pixel coordinates of the pixel point on the image and the image center point respectively. The generated point cloud contains a large amount of information and can express the spatial distribution and surface characteristics of the target, as shown in Fig. 2(c).

2.2 Projection of Point Clouds

It is obvious that projecting the spatial point cloud onto three orthogonal directions can express the structure of gesture and its changing regularity more clearly. In essence, projection from point cloud to plane can be regarded as the projection from three-dimensional space to two-dimensional plane.

Plane Π can be determined by a pair of orthogonal vectors A=(e_1,e_2). L is the line-segment from origin of three-dimensional sensing coordinate system to Π , whose length is l. The intersection of L and Π is C. θ and φ represent the angle between L and X-Axis, Y-Axis respectively. Plane Π is determined as follows:

$$e_1 = [-sin(\theta), cos(\theta), 0]^T, \tag{4}$$

$$e_2 = [cos(\theta) \times sin(\varphi), sin(\theta) \times sin(\varphi), cos(\varphi)]^T. \tag{5}$$

Then, the projection of vector $\boldsymbol{m}$ could be calculated by the following formula:

$$\boldsymbol{p} = \boldsymbol{m} \cdot P, \tag{6}$$

where $\boldsymbol{p}$ is the projection of $\boldsymbol{m}$, P represents projection matrix. $\boldsymbol{p}$ can be represented by orthogonal vectors (e_1,e_2):

$$\boldsymbol{p} = x_1 e_1 + x_1 e_2 = Ax, \tag{7}$$

where $A^T(\boldsymbol{m} - \boldsymbol{p}) = \boldsymbol{0}$, hence the projection matrix is:

$$P = A(A^T A^{-1})A^T. \tag{8}$$

θ and ϕ are set to $-\pi/2$, 0 and 0, $\pi/2$ to get the side view and the vertical view of point clouds. The side view and vertical view of the reconstructed gesture are shown in Fig. 2(e) and Fig. 2(f). In order to improve the accuracy of the whole recognition task and the applicability to the scene, the frame difference sequence *D(x,y)* is used as the front view of gesture.

$$D(x,y) = |f_n(x,y) - f_{n-1}(x,y)|, \tag{9}$$

where $f_n(x,y)$ represents the n-th frame image.

3. THE COGNITIVE METHOD BASED ON RECONSTRUCTED MULTIPLE VIEWS

Video-based gesture recognition involves not only the spatial characteristics of gesture appearance but also the temporal characteristics of movement changes. Besides, the fusion of three view features can achieve robust and stable spatiotemporal feature learning. In this section, MV3D-CNN is proposed to capture the signicant features of gesture in both temporal and spatial dimensions, as shown in Figure 3.

3.1 3D Convolution Neural Network

3D Convolution Neural Network is developed on the basis of 2D CNN. In order to thoroughly integrate movement information into gesture classification, it regards the time dimension of video sequence as the third dimension. Performing 3D convolution and 3D pooling instead of 2D operations in convolution and pooling stages, so as to cbtain the effective features of space-temporal dimension simultaneously.

3D convolution kernel connects the feature map with adjacent frames of the previous layer to perceive movement information. In convolution operation, an element *X(i,j,k)* of the n-th feature map in *num*-th layer can be expressed as:

$$X_{num,n}^{(i,j,k)} = g(W_n + b_{num,n}), \tag{10}$$

$$W_n = \sum_{N=1}^{N_{num-1}} \sum_{t=0}^{f_T} \sum_{w=0}^{f_W} \sum_{h=0}^{f_H} \varpi_{num,n,N}^{(x,y,z)} \cdot X_{num-1,N}^{(i+t,j+\varpi,k+h)} \tag{11}$$

where, *g(·)* represents activation function, which is used to improve the expression ability of neural network. f_T, f_W, f_H represent the depth, width and height of 3D convolution kernel. N_{num-1} represents the number of convolution kernels of the previous convolution layer, and ϖ represents the element value corresponding to the positions *(x,y,z)* of the n-th feature map in *num*-th layer and the N-th feature map of the *(num-1)*-th layer in the 3D convolution kernels. $b_{num,n}$ is the bias of the n-th feature map in *num*-th layer.

3.2 Architecture and Optimization of MV3D-CNN

The consecutive gesture has temporal-spatial property, which is difficult for 2D CNN to capture. Therefore, we propose MV3D-CNN to preserve the continuity of gesture from three perspectives in terms of the time dimension. Gesture features have temporal variability. On the premise of the broadest equal interval, 16 frames are selected from a complete video sequence as the input of the model. The gesture sequences of the three views obtained by the point cloud projection are input to front-view CNN, side-view CNN and vertical-view CNN, respectively.

Convolution layer and subsampled layer alternately extract features to train the data in an end-to-end way and then output the relative probability between different categories through the classification network. The framework of MV3D-CNN is shown in Figure 3.

MV3D-CNN has three input terminals, which input gesture sequences from three orthogonal perspectives, respectively. The size of each frame is 320×240 in the original video. Without loss of generality, we clip the input image of the front view to 192×144 in this paper. We only retain the gesture area of interest through the depth threshold

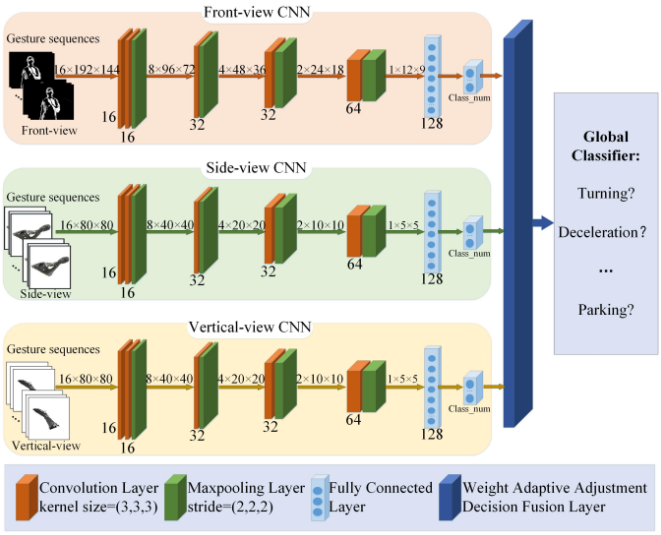

Fig. 3. MV3D-CNN: Convolutional network architecture of three orthogonal views. The network includes convolution layer, pooling layer and fully connected layer. The network architecture and architecture parameters are shown.

strategy in the side and vertical view. Therefore we cut the side and vertical view into 80×80 uniformly.

Five convolution operations alternate with four pooling operations. To reserve original image information and shorten the length of the temporal-spatial sequence, the first convolution layer includes two continuous convolution operations.

Convolution kernel size of each layer is 3×3×3 and the stride size is 1×1×1. The scale of pooling is set to 2×2×2 as well as pooling pattern is maximum pooling. Then the feature map passes through the full connection layer and softmax layer to get the output of a single view. Finally, the final classification score is obtained by decision fusion, which will be discussed in the next section.

For a neurons of convolution layer, the more connections it has with the neurons in the upper layer, the more information of original image is perceived. Therefore, these neurons may contain more semantic information and global features. In contrast, smaller receptive fields contain more local and detailed information. Based on this intuition, we consider optimizing the structure of MV3D-CNN to save calculating cost while keeping the receptive field unchanged.

In order to distinguish, we call the above model as MV3D-CNN$_a$, and the optimized model as MV3D-CNN$_b$. For MV3D-CNN$_b$, in the first layer of each network branch, we use 5×5×5 convolution kernel to replace two 3×3×3 convolution kernels, which can ensure that the size of receptive field will not change for a neuron. The change of convolution kernel results in the compression of architecture and the reduction of network depth and the saving of computing resources. Then in next layers, the convolution kernel whose size is 3×3×3 is split into 1×1×3, 1×3×1 and 3×1×1. The combination of three new convolution kernels can replace the original convolution kernels, produce the same receptive field and reduce the number of parameters(k^3 to $3k$).

3.3 Weight Adaptive Ajustment Decision Fusion

Effective combination of multiple classifiers can improve the accuracy and generalization of classification. For the parallel combination model of multiple classifiers, the weight adaptive adjustment decision fusion(WAADF) method is adopted to match the corresponding weight according to the decision result of each member classifier to form the final decision.

Assuming that N member classifiers are used to identify C types of gesture sequences, and D_n(n=1,2... N) represents the output decision of each classifier. Softmax regression outputs the posterior probability of feature f belongs to each category:

$$D_n(f) = (p_n(Ges_1|f), p_n(Ges_2|f), ..., p_n(Ges_C|f)), \quad (12)$$

$$\sum_{i=1}^{C} p_n(Ges_i|f) = 1. \quad (13)$$

Global classifier obtains better decision performance than each member classifier by fusing effective decision. WAADF dynamically integrates the decision information of each member classifier to achieve global classification. The difference of posterior probability of each category output by a member classifier is the primary basis for determining the weight proportion. For a member classifier i, the decision difference is obtained by comparing the posterior probabilities of all categories:

$$S(i) = \sum_{j=1}^{C-1} \sum_{k=1}^{C} |p_i(Ges_j|f) - p_i(Ges_k|f)|. \quad (14)$$

The weight obtained by classifier i in the global classification decision fusion process are:

$$W(i) = \frac{S(i)}{\sum_{j=1}^{N} S(j)}. \quad (15)$$

The more significant the decision difference of member classifier is, the higher the decision reliability is, and the larger the value of adaptive weight is. The decision of global classifier is as follows:

$$p_{global}(Ges_i|f) = \sum_{n=1}^{N} W(n) \cdot p_n(Ges_i|f). \quad (16)$$

The result is used to choose the gesture type with the most substantial posterior probability as the final decision.

4. DESIGN AND ANALYSIS OF EXPERIMENTS

For the network training, we use small batch random gradient descent algorithm to train the model. Considering the GPU factor, we use 64 samples to update the parameters in each iteration, which can reduce the fluctuation of the parameters update, ultimately get more stable convergence.

The fundamental purpose of this section is to evaluate proposed MV3D-CNN on IsoGD and then we compare

against the state-of-the-art algorithms for the task of gesture recognition. First of all, IsoGD and the training details of MV3D-CNN are introduced in section 4.1. In section 4.2, we describe the empirical analysis. We compare multiple network frameworks to verify the effectiveness of the proposed network in section 4.3.

4.1 Dataset and Training Details

IsoGD(see Wan et al. (2016)) is a vast isolated gesture data set sorted and published based on Chalearn gesture data set. We randomly selected 32,000 samples as training data. The input of each branch of MV3D-CNN is fixed to 16 frames by the maximum interval sampling, and then the image is clipped to 192×144(front view), 80×80(side view) and 80×80(vertical view). Considering the randomness of gesture and the uncertainty of shooting environment, we expand the data set through translation and random cutting to realize data enhancement: translate the data up 10 pixels, translate the data right 10 pixels, rotate clockwise by 5 degee and 10 degee.

The initial learning rate is set to 0.001, dividing the learning rate by 5 every 1000 iterations, and we set the total number of iterations to 100000. Finally, the experiments are performed on a single NVIDIA GeForce GTX TITAN X (8GB) GPU equipped with Intel(R) Core(TM) i5-8650K CPU @ 3.00GHz and 16GB RAM.

4.2 Results Analysis

Experiments evaluate our MV3D-CNN architecture. As shown in Fig. 4, the effects of multi-view fusion strategy and three single view fusion strategies are compared. Compared with the multi-classifier mean fusion strategy and the maximum fusion strategy, Fig. 5 shows the advantages of WAADF.

As shown in Fig. 4, the accuracy of using only vertical view and side view as training data are 38.23% and 47.88%, respectively. The results demonstrate that the vertical and side view can supplement three-dimensional contour of gesture, but the lack of grasp of the overall movement direction of a dynamic gesture, the specificity of different gesture is not apparent.

Compared with the above two views, front view of gesture sequence contains more motion information, and the specificity between different categories is evident, which can obtain higher accuracy of recognition task. Combination of the front view, side view and vertical view can master the three-dimensional features of gesture contour and hand movement pattern at the same time, so the recognition accuracy of single view for gesture sequence is not as high as MV3D-CNN. Due to the different number of network parameters, it can be observed that the convergence rate of side-view CNN and vertical-view CNN is faster than front-view only. Furthermore, it can be seen from Fig. 4 that when the branch neural network tends to be stable, MV3D-CNN also begins to converge.

The maximum decision fusion(MAX-DF) algorithm selects the maximum relative probability of member classifiers as global decision result while the mean decision fusion(MEAN-DF) algorithm averages the output of each

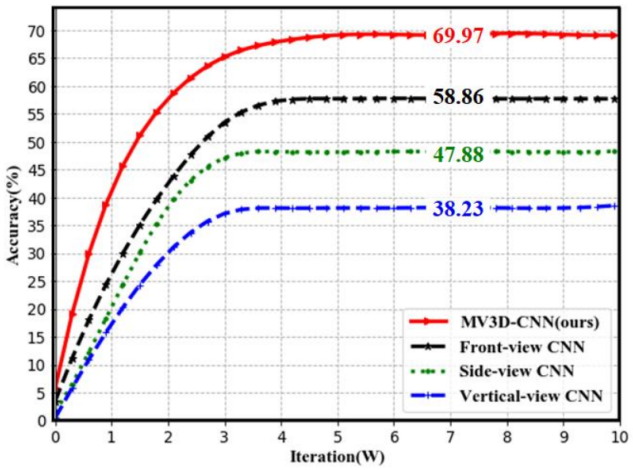

Fig. 4. The comparison between the front view, side view and vertical view separately with three views combination.

classifier as a global decision result. As shown in Figure 5(a), WAADF outperforms these two fusion algorithms by 7.83% and 15.14% in MV3D-CNN$_b$ architecture, while 7.87% and 14.83% in MV3D-CNN$_a$ architecture. It is a potent indication that WAADF strategy has played a positive role in improving network performance and decision making accuracy. We also tested each form of the framework we proposed. As shown in Fig. 5, the convergence rate of MV3D-CNN$_a$ is significantly slower than that of MV3D-CNN$_b$ when using the same decision fusion method as the number of parameters of the optimization model decreases.

4.3 Comparison

We compare the model proposed in this paper with the published methods for gesture recognition on IsoGD, and the comparative result as shown in Table 1. With the help of C3D model, large-scale gesture recognition method based on RGB-D videos(LSGR-RGBD) achieves 54.50% accuracy by fusing motion features and spatiotemporal features, see Zhu et al. (2017). 8-MFFs-3flc divides the gesture video into eight segments. By adding three optical ow frames to an RGB image, the optical flow information expressing motion is fused into a static image. The accuracy of 57.40% is achieved on IsoGD, see Kopuklu et al. (2018). 2SCVN-3DDSN, the unified framework of two-stream consensus voting network aggregating 3D depth salience convent stream simulates the short-term and long-term structure of RGB sequence, and the recognition accuracy reaches 67.19%, see, Duan et al. (2016). In contrast,

Table 1. Results and Comparison

Method	Result
LSGR-RGBD	54.50%
8-MFFs-3flc	57.40%
2SCVN-3DDSN	67.19%
MV3D-CNN$_a$ with MAX-DF(ours)	55.02%
MV3D-CNN$_a$ with MEAN-DF(ours)	61.98%
MV3D-CNN$_a$ with WAADF(ours)	69.85%
MV3D-CNN$_b$ with MAX-DF(ours)	54.83%
MV3D-CNN$_b$ with MEAN-DF(ours)	62.14%
MV3D-CNN$_b$ with WAADF(ours)	**69.97%**

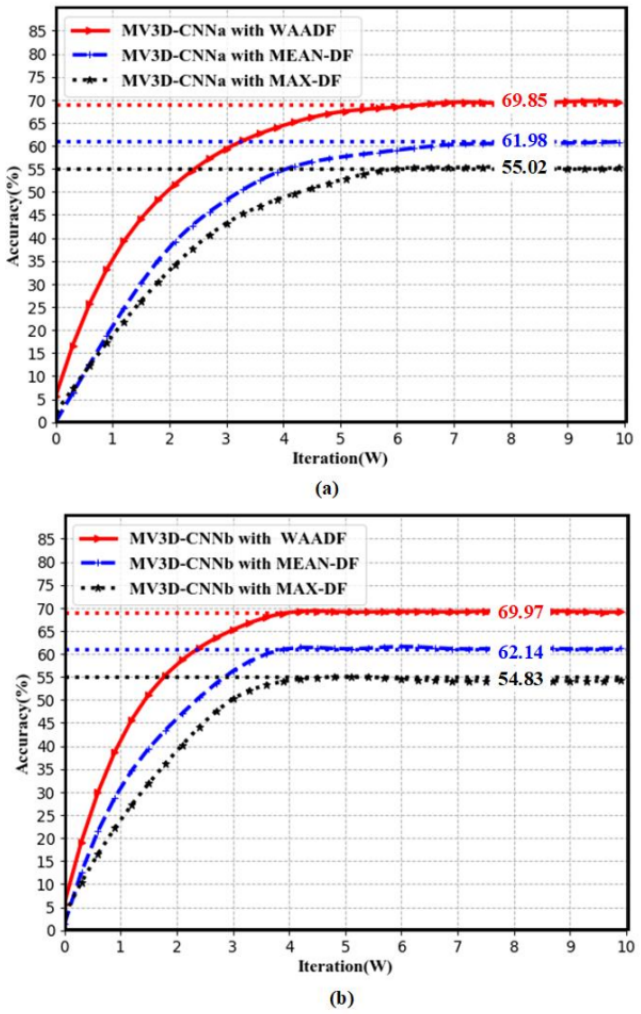

Fig. 5. Results of proposed frameworks with different decision fusion strategies.

the network architecture proposed in this paper has better performance. It is consistent with our observation because the accuracy of video recognition depends on the understanding of the whole sequence.

5. CONCLUSION

In this paper, an end-to-end deep learning framework based on three orthogonal views is proposed for gesture recognition of high-speed train drivers. Specifically, we first transform the RGB-D data format of gesture sequences into three-dimensional point clouds, by projecting which the side and vertical views of gesture are obtained. Then, an architecture named MV3D-CNN is designed and trained to learn the spatiotemporal movement of gestures.The numerical results verify the effectiveness of our solution and demonstrate the superiority of MV3D-CNN, compared to some state-of-the-art techniques.

6. ACKNOWLEDGEMENTS

This work is partially supported by National Natural Science Foundation of China (Grant Nos. 61873353 and 61672539).

REFERENCES

Asadi-Aghbolaghi, M., Claps, A., Bellantonio, M., Escalante, H.J., Ponce-Lpez, V., Bar, X., Guyon, I., Kasaei, S., and Escalera, S. (2017). A survey on deep learning based approaches for action and gesture recognition in image sequences. *12th IEEE International Conference on Automatic Face Gesture Recognition*, 476–483.

Da Gama, A., Fallavollita, P., Chaves, T., Silva Figueiredo, L., Baltar, A., Ma, M., Navab, N., and Teichrieb, V. (2016). Mirrarbilitation: A clinically-related gesture recognition interactive tool for an ar rehabilitation system. *Computer Methods and Programs in Biomedicine*, 135, 105–114.

Duan, J., Zhou, S., Wan, J., Guo, X., and Li, S. (2016). Multi-modality fusion based on consensus-voting and 3d convolution for isolated gesture recognition.

Gustin, F., Rendulic, I., Miskovic, N., and Vukic, Z. (2016). Hand gesture recognition from multibeam sonar imagery *. *IFAC-PapersOnLine*, 49(23), 470–475.

Kopuklu, O., Kse, N., and Rigoll, G. (2018). Motion fused frames: Data level fusion strategy for hand gesture recognition.

Molchanov, P., Gupta, S., Kim, K., and Kautz, J. (2015). Hand gesture recognition with 3d convolutional neural networks. In *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Plouffe, G. and Cretu, A.M. (2015). Static and dynamic hand gesture recognition in depth data using dynamic time warping. *IEEE Transactions on Instrumentation and Measurement*, 65(2), 305–316.

Pramod, Kumar, Pisharady, Martin, and Saerbeck (2015). Recent methods and databases in vision-based hand gesture recognition: A review. *Computer Vision and Image Understanding*, 141, 152–165.

Rautaray, S.S. and Agrawal, A. (2015). Vision based hand gesture recognition for human computer interaction: a survey. *Artificial Intelligence Review*, 43(1), 1–54.

Sharma, K., Moon, I., and Kim, S.G. (2012). Depth estimation of features in video frames with improved feature matching technique using kinect sensor. *Optical Engineering*, 51(10), 1–11.

Tran, D., Bourdev, L., Fergus, R., Torresani, L., and Paluri, M. (2015). Learning spatiotemporal features with 3d convolutional networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 4489–4497.

Wan, J., Li, S., Zhao, Y., Zhou, S., Guyon, I., and Escalera, S. (2016). Chalearn looking at people rgb-d isolated and continuous datasets for gesture recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 761–769.

Wang, Y., Zhao, S.G., Wang, J., and Wang, K. (2012). Research on monitoring plan of fatigue driving status for locomotive driver. *Railway Computer Application*, 12, 7–10.

Zhu, G., Zhang, L., Shen, P., and Song, J. (2017). Multimodal gesture recognition using 3-d convolution and convolutional lstm. *IEEE Access*, 5, 4517–4524.