Multi-channel Features Fitted 3D CNNs and LSTMs for Human Activity Recognition

Yang Qin, Lingfei Mo*, Jing Ye, Zhening Du
School of Instrument Science and Engineering
Southeast University
Nanjing, China
lfmo@seu.edu.cn

Abstract—**human activity recognition has been widely used in many fields, especially in video surveillance and virtual reality, etc. The paper investigates a general feature combination method for a relatively new 3D CNNs and LSTMs fusion model in human activity recognition. All the features used in this combination method are from human activity videos without manually extracting features or any prior knowledge, and the model has good generalization performance. Through extracting multi-channel features of the motion optical flow vector, grayscale and body edge, putting them to 3D convolutional neural network, and processing time characteristics within Long-Short Term Memory neural network, the recognition rate of the model rises greatly. The experiment selects KTH dataset as the data source. The model based on RGB is used to compare with the model based on multi-channel features. It shows that multi-channel features can improve recognition accuracy rate obviously, and have great robustness in different scenes, which proves that it is an efficient feature combination method fitted 3D CNNs and LSTMs.**

Keywords—***Human Activity Recognition; multi-channel features; 3D convolutional neural network; Long-Short Term Memory neural network.***

I. Introduction

With the spread of video recording technique, like camera, recognizing human activity from video sequences has been a widely studied topic, which is the most suitable recognition method for industrial practice.

A difficult point in human activity is that how to describe activities effectively. Underlying visual information is supposed to be screened in order to filter invalid features, but for the different performance of the same action, it is difficult to find a general feature description method to completely record human activity. Researchers try to describe activities by modeling human body. For example, the earliest 12 points human body model proposed by Johansson [1]. Davis stored motion information by two modules, MEI and MHI [2]. Mori and Malik described silhouette shape-by-shape context descriptor to get the 3D human posture information [3]. Hsuan-Shen extracted silhouette information, which used star skeleton to describe the angle between human baselines [4]. Batra saved the silhouette of STV and sampled STV with small 3D binary space blocks [5]. Laptev extended Harris corner to 3D Harris, which can describe local features efficiently [6].

It cannot be denied that human body models indeed describe the correlation properties of motions as far as possible. However, extracting a human body model from a video can only be applied in the situation of single and fixed background because of the difference of background in the original video and the effect of the temporal change. It is hard to extract the human body model when the environment or the visual angle changes.

The solution of this difficult problem is to extract features automatically by recognition algorithms. In this way, feature selection does not depend on the particular scene of the task, which means model parameters have transferability. The concept of automatic feature extraction has been used in deep learning. Deep network understands human activities by abstracting video images layer by layer and is of good generalization.

3D convolutional neural network and Long-Short Term Memory neural network are both used in human activity recognition. For instance, Ji et al. and Jhuang et al. classified KTH activity videos through the 3D convolutional neural network [7, 18]. Jeff Donahue et al. combined 2D convolutional neural network and Long-Short Term Memory neural network and proposed Long-term Recurrent Convolution Network [8]. Baccoucheet et al. combined 3D convolutional neural network and Long-Short Term Memory neural network and achieved good recognition results [9].

In this paper, based on the fusion model of the 3D convolutional neural network and the Long-Short Term Memory neural network proposed by Baccouche, a general feature combination method is proposed. The model receives three-channel features, the motion optical flow vector, gray scale and the body edge image. Moreover, this model is capable of learning human activities gradually.

KTH dataset is the data source of the experiment. The background of KTH dataset is single and the environment is simple, so the motions are of high discriminability. In addition, it has four different scenes. The effects of environment are compared according to the training situation in different scenes. The paper chooses this data set in order to prove the robustness of feature combination method on the external environment.

The experiment result suggests that the recognition rate of

978-1-5090-0795-0/15/$31.00 ©2016

the model increases, especially in the data set with changing environment, which means that the feature combination method is fit to the fusion model of 3D CNNs and LSTMs.

II. MULTI-CHANNEL FEATURES

If a single feature is used, the recognition efficiency is likely to change substantially when task scene changes. Thus, for an action, a variety of features that are independent of each other are used, and features can complement each other.

A. *Motion Optical Flow Vector*

Motion optical flow vector is used to describe motion coherence[12].As Fig.1 shown, computing the instantaneous velocity field caused by the movement of the pixel on spatial moving objects, the motion optical flow vector is obtained from the luminance information of two continuous frames in image sequences [13]. The motion optical flow vector is an independently feature without any knowledge of human body and shape information.

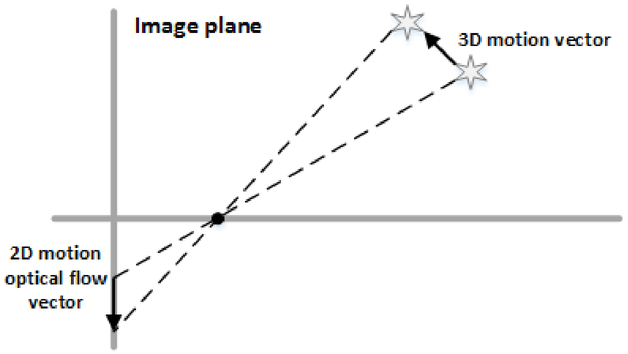

Fig. 1. Optical flow field is the projection of the motion field in the 2D image plane.

The optical flow calculation is easily influenced by the noises. Therefore, other features, which are not sensitive to noises, must be considered together. .

Certainly, optical flow method can test separated moving targets with a moving camera, which is the advantage of the motion optical flow vector and can supply the features that are sensitive to background movement.

B. *Grey Scale*

For human activity recognition, the relevance between human body and environment is also useful. With convolutional neural network, the original images can be directly abstracted. Doing gray processing can maintain the overall appearance information and filter invalid color factor.

Gray scale is not sensitive to noise because only substantial original image information needs to be preserved, which means it can compensate for other features which are sensitive to noise

C. *Body Edge Image*

The time-varying characteristic of body shape is used to classify activities, because the model has Long-Short Term Memory neural network, which can extract information of time series efficiently.

To study the change of body shape, the human body have to be separated from the environment by edge detection.

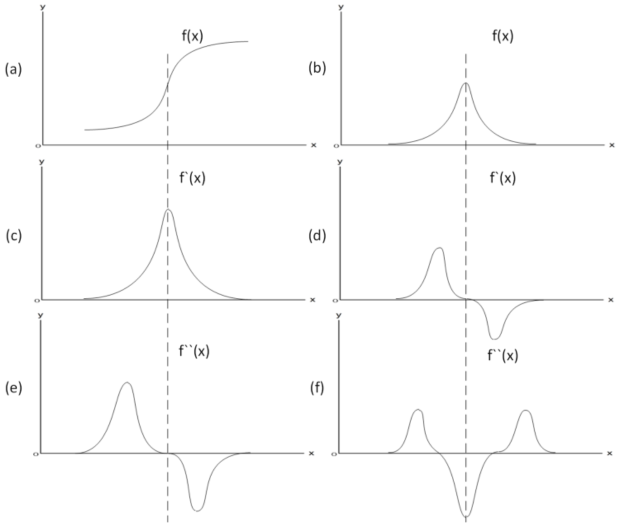

Fig. 2. (a)(b)are the edge changed in the form of step function and the edge changed in the form of rooftop function, respectively; (c)(d)are their first derivatives; (e)(f)are their second derivatives.

Fig.2 shows different types of edge[14]. The algorithm of Canny is used to adapt to different edges: first, does Gauss filter to images, and then does weighted average in certain parameter principle according to pixels. Finite difference of first partial derivatives is used to calculate the amplitude and direction of gradients. Then, non-maximum suppression is used on gradient amplitude. Global gradients cannot determine edges. The local maximum of pixels have to be found. Set the gray value of non-maximum points as zero and eliminate a majority of non-edge points. Finally, the paper uses dual threshold algorithm and connecting edge.

Similar with the motion optical flow vector, body edge images have bad performance on anti-noises. Excessive filtering will lead to a fuzzy boundary, which must be complemented by gray features.

III. MODEL STRUCTURE

In this chapter, model internal structure will be described. Parameters have been given to repeat the experiment.

A. *System Work flow*

Fig.3 shows the system workflow. Sample the video by a certain frame rate and do pre-processing on the images in the sequence. Labels ought to be added to data in advance. Through training data, model adjusts parameters and uses testing data to examine.

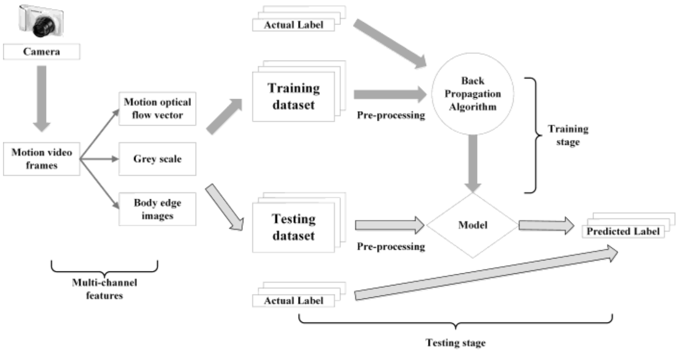

Fig. 3.System workflow.

B. Deep Neural Network

1) 3D Convolutional Neural Network.

3D Convolutional Neural Network is stacked by alternating convolution layers and sub-sampling layers[16, 17]. Between each convolution layer and sub-sampling layer, there is an activation layer to accelerate the convergence and propagate gradients as far as possible. Rectified Linear Units (ReLU) are used to replace sigmoid function to activate neuron.

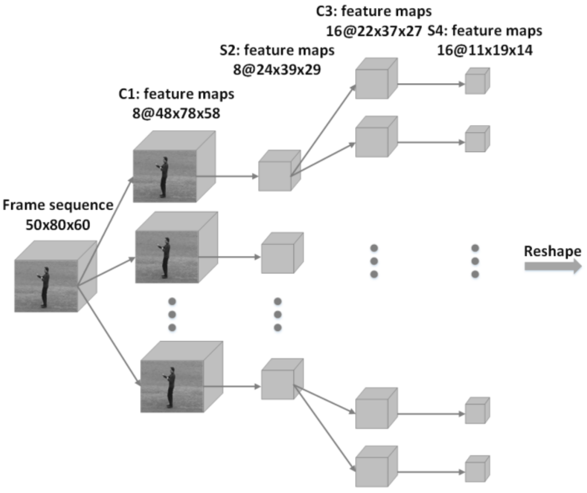

Fig. 4. 3D Convolutional Neural Network

The model is formed according to the method in reference paper [18]. The convolution kernel size is designed as (3*3*3) and the pooling size as (2*2*2). The input data is an image sequence of 50 frames. The data size is 50*60*80. As Fig.4 shows, C1 layer has 8 different 3D convolution kernels. The size of feature map is 48*58*78. S1 layer does max pooling on the variables from C1 layer. The pooling result lowers the input dimensions largely, and the data size changes into 8*24*29*39. C2 layer has 16 different 3D convolution kernels. The data size changes into 16*22*27*37 after convolution. S2 layer receives the input from C2. After a sub-sampling with the pooling size of 2*2*2, the final output variable size changes into 16*11*14*19=46816.

2) Reshape Layers.

A reshape layer is connected after the 3D convolutional neural network to adjust data forms. As Fig.5 shows, the essence of reshape layer is a special fully connected network. The output of 3D convolutional neural network is regrouped as 2D array first, then the features of each time slice are put into fully connected network to lower more feature size.

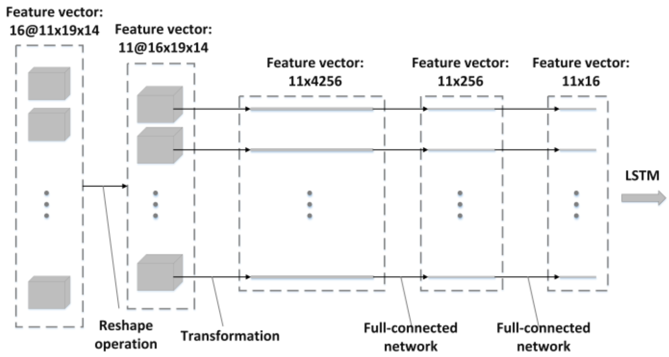

Fig. 5. Reshape layers change the size of features

3) Long-short Term Memory Neural Network.

The feature sequence that contains temporal information is processed by long-short memory neural network [19].The network of LSTM is shown in Fig.6. A merge layer is connected behind LSTM layer. The merge layer does mean treatment to LSTM output values at all moments.

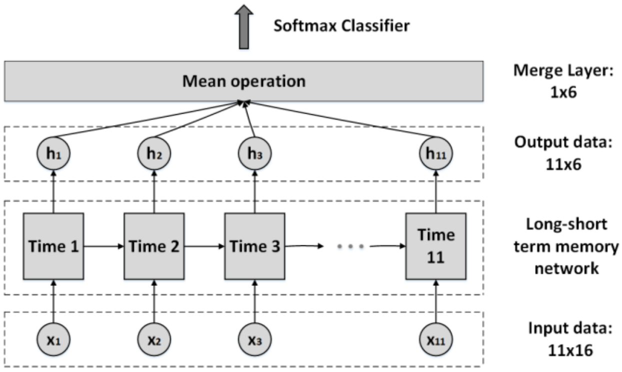

Fig. 6. Long-short term memory neural network

3) Network Architecture.

Fig.7 shows the overall framework of the fusion model. The model uses softmax classifier, so the labels are one-hot encoding, only setting the corresponding bit as one.

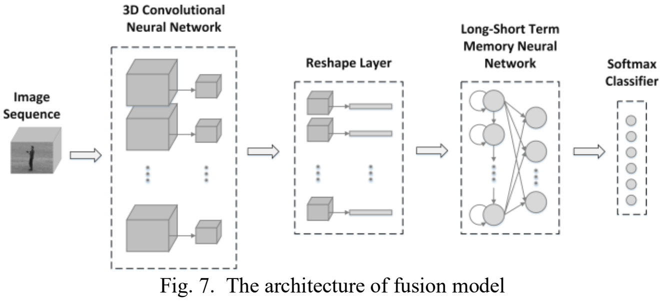
3D Convolutional Neural Network
Image Sequence
Reshape Layer
Long-Short Term Memory Neural Network
Softmax Classifier

Fig. 7. The architecture of fusion model

C. Dropout Tricks

The work principle of Dropout is to improve overfitting by preventing co-adaptation of feature detectors[20]. As is shown in Fig.8, randomly let certain weights of hidden layer nodes in the network stop working, and temporarily regard them as a part beyond the network, and their weights are stored.

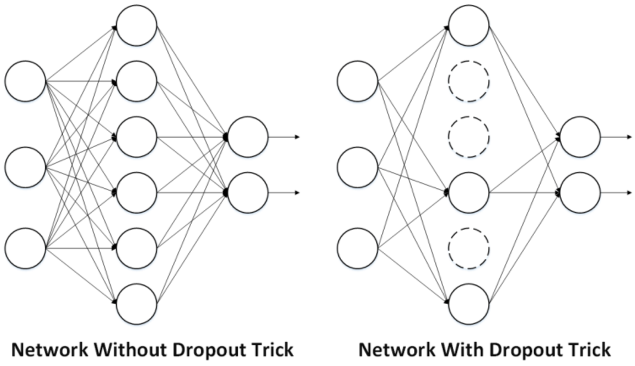

Fig. 8. Dropout will randomly forbid half of the nodes to avoid updating.

IV. EXPERIMENT AND RESULTS

A. KTH Description

KTH dataset, which is shown in Fig.9, consists of six types of human actions (walking, jogging, running, boxing, hand waving and hand clapping). These actions were performed by 25 people in four different scenes. All sequences were taken over homogeneous backgrounds with 25fps frame rate. The sequences were down sampled to the spatial resolution of 160×120 pixels.

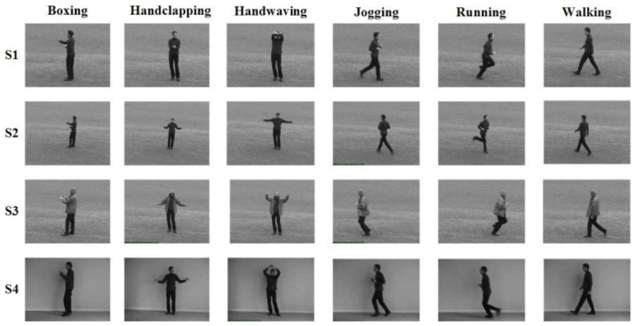

Fig. 9. KTH data set contains 6 actions which were performed by 25 people in four different scenes: outdoors s1, outdoors with scale variation s2, outdoors in different clothes s3 and indoor s4.

The way to process dataset is shown below:

- The images in the frame sequence were compressed to 80×60 pixels for a smaller calculation size;
- The video was sampled by 50 fps and marked as a sample every 50 frames;
- 60% of the samples were as training data and 40% of the samples were as testing data;

B. Data Processing

1) Gaussian Filtering.

Gaussian filtering can remove isolated noise and store edge features of human body, which will not fuzz the image obviously and avoid information loss in the processing of canny edge detection operator [15].

2) Standardizing Input Data.

Each pixel value is in the range of [0,255]. It will lead to a low learning efficiency of the model because sigmoidal neurons in the model are saturated quickly. Therefore, each pixel value need to be divided by 255 to make the data in the range of [0, 1]. Then process the data through subtracting each frame of the sequence by the mean of all frames in a sequence, and center each pixel value in the range of [-0.5, 0.5].

3) Shuffling Samples.

Irrational sample distribution will probably cause fast convergence on parameters and involve in local optimal solution. The data labels should be dispersed as far as possible and adjacent samples come from different type. Random samples are better than sequential samples on improving the generalization performance. Stochastic algorithms are used to create disordered index to recombine data.

4) K-fold Cross Validation

To eliminate differences such as age or weight, the division is not in accordance with people. Combine the whole sample set and divide it into K subsets with equal number of samples. Use (K-M) subsets as the training data set, M subset as the test data set in each experiment. The experiment sets K=5 and M=2. Traverse all combinations and select the best result as outcome.

C. Framework

The paper uses python to build the model with Tensor Flow. The environment of experiment contains NVIDIA GTX 960 1024 CUDA cores / Intel i3 2.26GHz / 8GB DDR3 Memory / Ubuntu 14.01 x64.

D. Results and Analysis

Divide data set into five subsets. To meet the requirement of supervised learning, training data set contains three subsets and testing data set contains two subsets.

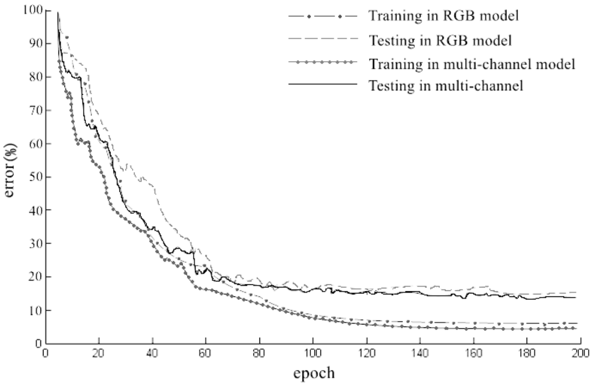

Fig. 10. With the increasing of epoch, error gets smaller and accuracy rate is increased and stable.

The horizontal axis of Fig.10 represents training epoch. The vertical axis represents error rate of models. With the increase of epoch, the error rate reduces. Compared with model based on RGB, Model based on multi-channel features can generate better recognition rate.

Table I and Table II use confusion matrix to visualize the training result. The average recognition rate of the model based on RGB information is 90.8%, while using multi-channel feature can increase the average recognition rate to 94.3%.

TABLE I. RGB MODEL RESULT BY CONFUSION MATRIX

		Predicted					
		Box	Jog	Wav	Clap	Walk	Run
Actual class	Box	90	1	6	1	1	1
	Jog	-	86	-	-	5	9
	Wav	3	-	92	4	1	-
	Clap	-	-	2	96	1	1
	Walk	-	1	1	-	96	2
	Run	1	9	-	-	5	85

TABLE II. MULTI-CHANNEL MODEL RESULT BY CONFUSION MATRIX

		Predicted					
		Box	Jog	Wav	Clap	Walk	Run
Actual class	Box	93	4	1	1	-	1
	Jog	-	92	1	-	3	4
	Wav	1	-	95	3	-	1
	Clap	-	-	3	95	2	-
	Walk	-	1	1	-	98	-
	Run	-	2	-	1	4	93

To study multi-channel features in different environments, the experiment do individual training for each scene.

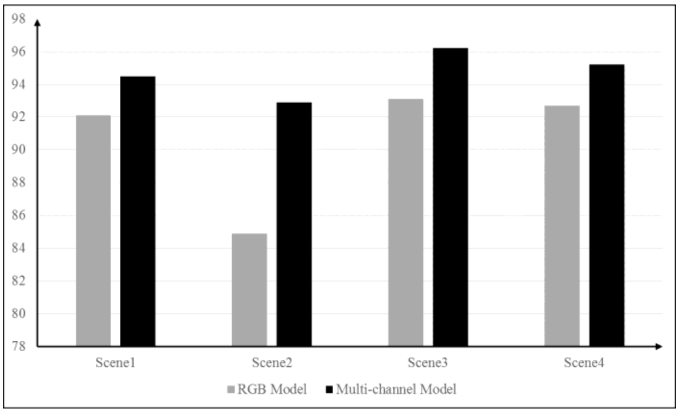

Fig. 11. For the scene with complex background, the improvement of multi-channel features on model recognition rate is more obvious.

In four scenes, only scene 2 changes the scale background. Fig.11 shows the recognition rates of RGB model and multi-channel features model. Multi-channel features improve the recognition rate in four scenes, and in the scene two, the effect is especially obvious.

Related research proposes some method for traditional model [21, 22]. However, these methods cannot make full use of 3D CNNs and LSTMs [7, 9]. Multi-channel features describe actions in three aspects, and show great performance in 3D CNNs and LSTMs fusion model.

V. CONCLUSION

In this paper, multi-channel features applying to 3D CNNs and LSTMs fusion model are proposed. In KTH dataset, compared to directly extracted features of RGB images, the model raises the recognition rate from 90.8% to 94.3% by learning multi-channel features. In addition, multi-channel features are of less background interference.

Although the experiment verified the advantages of multi-channel features, further detailed researches remain to be done. If some features, which are proved to be efficient, such as 3D-SHIFT and cuboids descriptor, apply to the model, the model can be improved [22]. For the future study, continue work on the features will be conducted.

ACKNOWLEDGMENT

The authors gratefully acknowledge the funding provided by the Fundamental Research Funds for the Central Universities under Grant 2242014R30021.

REFERENCES

[1] Johansson G,. Visual motion perception.[J]. Encyclopedia of Human Behavior, 1975, 232(6)(6): 76-88.

[2] BobickA F, Davis J W. The Recognition of Human Movement Using Temporal Templates[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2001, 23(3): 257-267.

[3] Mori G, Malik J. Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA[C]// null. IEEE Computer Society, 2003:134.

[4] Chen, H. S., H. T. Chen, et al. (2006). Human action recognition using star skeleton, ACM.

[5] Batra, D., T. Chen, et al. (2008). Space-time shapelets for action recognition, IEEE.

[6] Laptev, I. (2005). "On space-time interest points." International journal of computer vision 64(2): 107-123.

[7] Jhuang, H., Serre, T., Wolf, L., and Poggio, T. A biologically inspired system for action recognition. In ICCV, pp. 1–8, 2007.

[8] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description [M]// AB initto calculation of the structures and properties of molecules /. Elsevier,, 1988:85-91.

[9] Baccouche M, Mamalet F, Wolf C, et al. Sequential deep learning for human action recognition[M]//Human Behavior Understanding. Springer Berlin Heidelberg, 2011: 29-39.

[10] Turaga P, Chellappa R, Subrahmanian V S, et al. Machine Recognition of Human Activities: A Survey[J]. Circuits & Systems for Video Technology IEEE Transactions on, 2008, 18(11): 1473 - 1488.

[11] Poppe R. A survey on vision-based human action recognition[J]. Image & Vision Computing, 2010, 28(6): 976-990.

[12] Royden C S, Moore K D. Use of speed cues in the detection of moving objects by moving observers[J]. Vision Research, 2012, 59(2): 17-24.

[13] Zhu G, Xu C. Action Recognition in Broadcast Tennis Video Using Optical Flow and Support Vector Machine[C]// 18th International Conference on Pattern Recognition (ICPR 2006), 20-24 August 2006, Hong Kong, China. 2006:251-254.

[14] Canny J. A computational approach to edge detection [J]. Pattern Analysis & Machine Intelligence IEEE Transactions on, 1986, PAMI-8(6):679-698.

[15] Haddad R A, AkansuA N. A class of fast Gaussian binomial filters for speech and image processing[J]. IEEE Transactions on Signal Processing, 1991, 39(3): 723-727.

[16] Wang J, Lu J, Chen W, et al. Convolutional neural network for 3D object recognition based on RGB-D dataset[C]// Industrial Electronics and Applications. IEEE, 2015.

[17] Cronje F. Human action recognition with 3D convolutional neural networks[D]. University of Cape Town, 2015.

[18] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for automatic human action recognition: US, US 8345984 B2 [P]. 2013.

[19] Graves A. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8): 1735-80.

[20] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors [J]. arXiv preprint arXiv:1207.0580, 2012.

[21] Moeslund T B, Hilton A, Krüger V. A survey of advances in vision-based human motion capture and analysis[J]. Computer Vision & Image Understanding, 2006, 104(2): 90-126.

[22] Salmane H, Ruichek Y, Khoudour L. Object tracking using Harris corner points based optical flow propagation and Kalman filter[C]// Conference Record - IEEE Conference on Intelligent Transportation Systems. 2011:67-73.