Journal of Korea Multimedia Society Vol. 22, No. 12, December 2019(pp. 1457-1465)
https://doi.org/10.9717/kmms.2019.22.12.1457

Human Action Recognition Based on 3D Convolutional Neural Network from Hybrid Feature

Tingting Wu[†], Eung-Joo Lee[††]

ABSTRACT

3D convolution is to stack multiple consecutive frames to form a cube, and then apply the 3D convolution kernel in the cube. In this structure, each feature map of the convolutional layer is connected to multiple adjacent sequential frames in the previous layer, thus capturing the motion information. However, due to the changes of pedestrian posture, motion and position, the convolution at the same place is inappropriate, and when the 3D convolution kernel is convoluted in the time domain, only time domain features of three consecutive frames can be extracted, which is not a good enough to get action information. This paper proposes an action recognition method based on feature fusion of 3D convolutional neural network. Based on the VGG16 network model, sending a pre-acquired optical flow image for learning, then get the time domain features, and then the feature of the time domain is extracted from the features extracted by the 3D convolutional neural network. Finally, the behavior classification is done by the SVM classifier.

Key words: 3D Convolution, Time Domain Features, Feature Fusion, VGG16

1. INTRODUCTION

Human motion analysis, behavior classification and behavior recognition in video image sequence have been part of the research hotspots in the field of computer vision [1]. In computer vision, human action recognition involves pattern recognition, image processing, computer vision, artificial intelligence and other fields. It is commonly used in human-computer interaction, action capture analysis, video monitoring and safety, environmental control detection and prediction [2-3]. Human action recognition is mainly affected by individual differences, angle of view changes, camera movement and illumination angle [4], it is still a challenging subject to accurately identify and analyze human behavior in real scenes, so it is very important to develop a set of advanced action recognition algorithm. Research on how to extract effective features from the video is essential to solve the above problems and design more effective behavior recognition framework [5]. At present, methods of human behavior recognition are mainly divided into methods based on traditional behavior recognition and the methods based on meaningful learning.

The habitual action recognition method is mainly composed of two steps. The first step consists of extracting the features of the video image; the second step is used by the learning classifier to classify the features. In the actual scene, different behaviors have obvious differences in appearance and movement mode, so it is difficult to discern the appropriate features, and deep learning model can learn by sample characteristics, which have the advantage of better than customary action recog-

※ Corresponding Author : Eung-Joo Lee, Address: 428, Sinseon-ro, Nam-gu, Busan, Kroea, TEL : +82-51-629-1143, FAX : +82-, E-mail : ejlee@tu.ac.kr
Receipt date : May 29, 2019, Revision date : Dec. 24, 2019
Approval date : Dec. 26, 2019

[†] Dept. of Information Communication Engineering, Tongmyong University
(E-mail : jangneyong0829@hotmail.com)
[††] Dept. of Information Communication Engineering, Tongmyong University

nition method [6]. In recent years, with the huge success of profound learning method in image classification and target detection, people begin to use deep learning method in video behavior recognition.

Convolutional neural network (CNN) is a deep network model, which can learn features from the original data [7]. Research shows that their method achieves superior performance in visual target recognition task. CNN is mainly primarily in the classification and detection of two-dimensional images. It mainly decomposes video frame into multiple still images and uses the CNN network model to identify the actions of a single video frame. However, this method ignores the action information in the continuous frame video image. In order to effectively learn the space-time characteristics in video sequence, a 3D CNN network model has been proposed [8-9]. By replacing the 2D convolution kernel with 3D convolution kernel for the convolution operation, spatial information and time information can be obtained at the same time. The network model generates multiple information channels from continuous multi-frame video images, convolves them on each channel, understands them, and finally merges the information of all channels to obtain time-space feature. In addition, an auxiliary feature is introduced to the 3D CNN network, and a normalized 3D CNN model is proposed. By combining the output of some different frameworks, the performance of the 3D CNN model is further improved.

2. HUMAN ACTION RECOGNITION BASED ON 3D CNN

Traditional video frame image processing mainly uses 2D convolution for feature extraction. Then, when the behavior is recorded, there is a certain regularity between successive frames of human motion. Therefore, 2D convolution cannot extract features according to this feature. In order to effectively synthesize motion information, 3D convolution is proposed [10]. The biggest feature of 3D convolution is the ability to extract features between successive video frame data cubes, which capture feature information in both time and space dimensions, and the operation processes multiple frames at once. Therefore, the 3D convolution layer and the pooling 3D layer are used to speed up processing while processing the timing.

2.1 3D convolution layer

The 3D convolution process is a process of superimposing successive video frame images into a cube and then convoluted with a 3D convolution kernel cube. Since each map finally generated by the operation is obtained by convolution of a plurality of adjacent continuous video frames of the previous layer, the effect of motion information capture can be increased. The 3D convolutional layer and the pooling 3D layer appear in three dimensions to form a 3D convolutional network. Fig. 1 shows the process of 2D convolution and 3D convolution.

3D convolution is a data set formed by superimposing a plurality of consecutive frames through a three-dimensional convolution kernel. A plurality of consecutive frames sequentially pass through a convolution layer, and each feature map in the convolution layer is adjacent to multiple layers of the upper layer. Continuous frames are connected to obtain certain motion information [11], which can be expressed as:

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

Among them, v_{ijm}^{xyz} represents the result value of the j-th feature pixel (x,y,z) of the i-th layer, $\tanh(\cdot)$ is a hyperbolic tangent function, b_{ij} is the deviation of the j-th feature map of the i-th layer convolutional layer, m is the number of feature maps of the $(i-1)$-th layer, P_i, Q_i, and R_i are the spatial and temporal dimensions of the i-th layer 3D convolution kernel, w_{ijm}^{pqr} is the convolution ker-

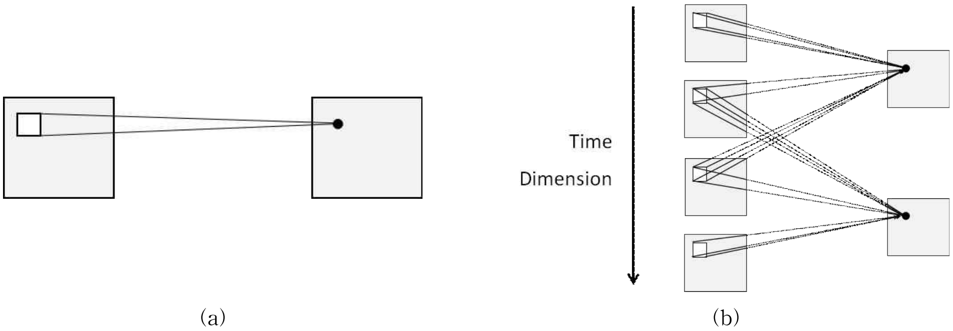

Fig. 1. 2D convolution(a) and 3D convolution(b).

nel weight of the m-th feature map connection of the previous layer.

2.2 3D pooling layer

After the video sequence passes through the 3D convolutional layer, a large number of image information features are acquired, so that the amount of data to be processed is greatly increased, and corresponding redundant data is generated accordingly, so the data needs to be down sampled [12]. Compared with the pooling 2D layer, the pooling 3D layer can simultaneously down sample data in both time and space dimensions, greatly reduce the size of feature maps, reduce redundant information, and reduce the connection between data, avoiding overfitting. Finally, to increase the classification accuracy. Similar to the pooling 2D layer, the commonly used 3D pooling layer sampling methods include maximum pooling, average pooling, and random pooling. The three-dimensional maximum pooling formula is as shown in equation (2):

$$v_{x,y,z} = \max_{0 \le i \le S_1, 0 \le j \le S_2, 0 \le k \le S_3,} (u_{x \times s+i, y \times t+j, z \times r+k}) \tag{2}$$

In the formula, the input vector of the pooling 3D layer is u, the output of the pooling process is v, and the sampling steps of the three directions are s, t, and r.

2.3 The deficiency of classical 3D CNN

When existing 3D CNN is convoluted in time, the 3D convolution kernel is used to convolute the continuous frame image cube in the form of a sliding window [13]. However, due to changes in pedestrian attitude, motion, and position, convolution at the same position is not appropriate, and when the 3D convolution kernel is convoluted on the timing, only the temporal characteristics of three consecutive frames can be extracted at a time, which is not very good to get sports information.

In order to improve the motion information, enrich the motion features and enhance the robustness of single feature representation, a behavior recognition method based on feature fusion of 3D convolutional neural network is proposed. In other words, based on the VGG16 network model, the pre-acquired optical flow images are sent for learning, and then the time-domain features are acquired. Finally, time-domain features are fused with the features extracted by 3D convolutional neural network for feature fusion, and the behavior classification is carried out by SVM classifier. Fig. 2 is active recognition structure based on feature fusion of 3D CNN:

3. TIME DOMAIN FEATURE EXTRACTION

In order to extract better time domain features, the model design of convolutional neural network is particularly critical. This paper is based on VGG16 network model [14], which was published in 2014. The network shows that stacking multiple layers is a key factor in improving computer vision

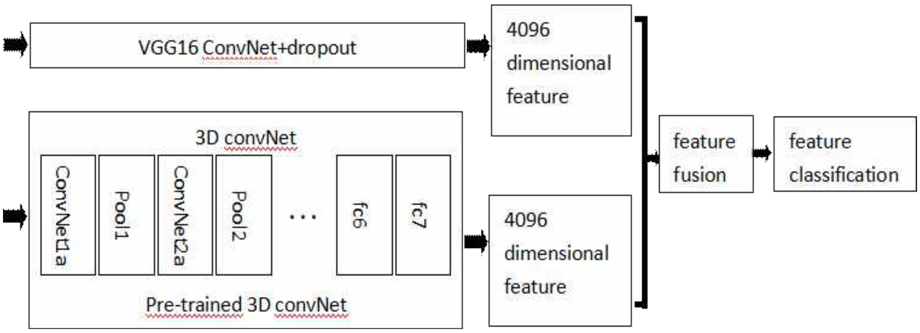

Fig. 2. Action recognition structure based on feature fusion of 3D CNN.

performance. The network features smaller size of the convolution kernel, smaller design of convolution step, smaller lower sampling window and deeper network structure. It is mainly composed of small 3×3 convolution operations and 2×2 pooling operation. In order to preserve the spatial resolution, edge processing is carried out in each convolution layer, that is, the size of the image is not changed during the convolution processing.

The advantage of VGG network is that stacking multiple small convolution kernels without using pooling operation can increase the representation depth of the network and limit the number of parameters at the same time. First, it combines three non-linear functions to make the decision function more perceptive and representational. Second, the parameters were reduced by 81%, while the receptive field remained unchanged. In addition, the effectiveness of different convolution kernels is improved.

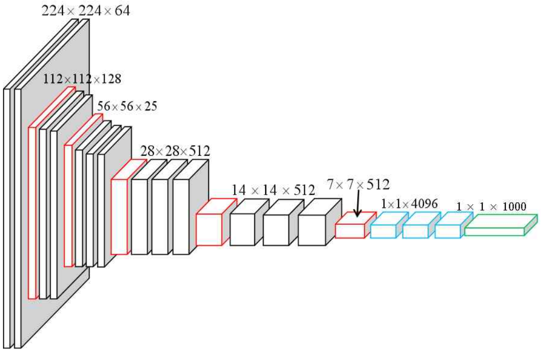

Fig. 3. Time domain convolution neural network.

The Fig. 3 is time domain convolutional neural network, it can be seen that the network model consist of 13 convolution layers, 5 pooling layers, 15 active layers, 3 fully connected layers, and 1 softmax layer. The size of the convolution kernel is 3×3, the sliding step of the convolution kernel is 1×1, and the edge processing is performed. The size of the sampling window of the pooling layer is 2×2. The Table 1 shows the output data size of data after processing in each network layer:

4. FEATURE FUSION AND FEATURE CLASSIFICATION

The fusion method is mainly divided into two aspects: feature fusion and result fusion.The methods of feature fusion are serial feature fusion, weighted feature fusion, and serial feature fusion or weighted feature fusion based on a series of feature correlation coefficients derived from the two fusion methods. Serial features fusion method enhances the robustness of single feature representation and achieves good recognition effect in the field of behavior recognition. At the same time, it has the advantages of simple fusion and simple calculation. However, serial feature fusion will increase feature dimension, which may lead to a large error in learning results.

In this experiment, we successively adopted serial feature fusion and weighted feature fusion.

Table 1. Network layer output data size

The network layer	Output data size	Number of filters	The network layer	Output data size	Number of filters
data	20×224×224				
conv1_1+relu1_1	64×224×224	64	conv1_2+relu1_2	20×224×224	64
pool1					
conv2_1+relu2_1	128×112×112	128	Conv2_2+relu2_2	128×112×112	128
pool2	128×56×56				
conv3_1+relu3_1	256×56×56	256	Conv3_2+relu3_2	256×56×56	256
conv3_3+relu3_3	256×56×56	256			
pool3	256×28×28				
conv4_1+relu4_1	512×28×28	512	Conv4_2+relu4_2	512×28×28	512
conv4_3+relu4_3	512×28×28	512			
pool4	512×14×14				
conv5_1+relu5_1	512×14×14		conv5_2+relu5_2	512×14×14	512
conv5_3+relu5_3	512×14×14				
pool5	512×7×7				
fc6+relu6	4096×1×1	4096	fc7+relu7	4096×1×1	4096
fc8	101×1×1	101			
loss	101×1×1				

(1) Serial feature fusion the method of serial feature fusion is to combine two sets of features in the sample space directly into a new feature vector, and then extract and compress the synthesized feature vector.

Suppose trainX and trainY are two different groups of characteristics, and the feature size of trainX is $M_1 \times N_1$, represented by $trainX_{M_1 \times N_1}$; and the feature size of trainY is $M_2 \times N_2$, represented by $trainY_{M_2 \times N_2}$. The combined feature $fusion_{concat} = [trainX, trainY]$ and requirements $M_1 = M_2 = M$. The combined feature size is $M \times (N_1 + N_2)$. In the experiment, the obtained 3D convolution feature size is 147926×4096 and the time-domain feature size is 147926×4096, so the feature size after direct serial feature fusion is 147926×8192.

(2) Weighted feature fusion weighted feature fusion method is to set different weights for the two groups of features in the sample space according to the proportion of feature participation, and then merge the features.

Suppose trainX and rainY are two different groups of characteristics, and the feature size of trainX is $M_1 \times N_1$, represented by $trainX_{M_1 \times N_1}$; and the feature size of trainY is $M_2 \times N_2$, represented by $trainY_{M_2 \times N_2}$. The weights of the two groups of features are w_1, w_2. Requirements $M_1 = M_2 = M$, $N_1 = N_2 = N$, and the combined feature $fusion_{weight} = [w_1 \times trainX + w_2 \times trainY]$. The combined feature size is $M \times N$. In the experiment, the obtained 3D convolution feature size is 147926×4096 and the time-domain feature size is 147926×4096, so directly weighted feature fusion feature size is 147926×4096.

In this experiment, two weight value settings are used: the first weight setting consideration is: C3D network feature extracts one feature for every 16 consecutive video images, and the VGG network feature extracts one feature for each successive 10 optical image, so Equation (1) get (2):

$$\begin{cases} w_1 + w_2 = 1 \\ \dfrac{w_1}{w_2} = \dfrac{16}{10} \end{cases} \tag{3}$$

$$\begin{cases} w_1 = 0.6154 \\ w_2 = 0.3846 \end{cases} \tag{4}$$

so $fusion_{weight1} = [0.6154 \times 3D + 0.3846 \times Vgg16]$.

The second weight setting consideration is to calculate the C3D feature and the average of the vector modes in the VGG network feature.

$$\begin{cases} w_1 \approx 0.6 \\ w_2 \approx 0.4 \end{cases} \tag{5}$$

so $fusion_{weight2} = [0.4 \times 3D + 0.36 \times Vgg16]$.

In this paper, SVM is adopted to classify the fusion features generated by each model. The UCF 101 data set contains 101 types of behaviors. Each type of behavior was performed by 25 different groups of people, and everyone has multiple video. Taking the first seven groups of each type of behavior as test samples, the last 18 groups were used as another test sample.

The test data are sent into the trained SVM classifier. Each classifier classifies and identifies the test data, It is considered to belong to the category that is classified into the most time. Finally, compared with the labels marked in advance, if the categories are consistent, the classification is considered to be correct.

$$Accuracy = \frac{correctly\,testing\,sample}{101 \times 18} \times 100\% \tag{6}$$

5. EXPERIMENT AND RESULTS

5.1 Image cropping of time domain feature extraction

Studies have shown that random cropping and horizontal flipping are very effective in preventing overfitting. Therefore, in the process of data learning and training of convolutional neural networks, an effective image enhancement technology is designed for network learning:

As showed in Fig. 4, an image is cropped into five images of size, and the image cropping is defined as the upper left image, the upper right image, the lower left image, and the lower right image. The image of the green border in the figure represents the image obtained by cropping in the upper left corner of the original image; the image of the blue border in the figure represents the image obtained by cropping in the upper right corner of the original image; the image of the red border in the figure indicates the cropping in the lower left corner of the original image. The image obtained afterwards; the image of the purple border in the figure represents the image obtained by cropping in the lower right corner of the original image; the image of the yellow border in the figure represents the image obtained by cropping in the center of the

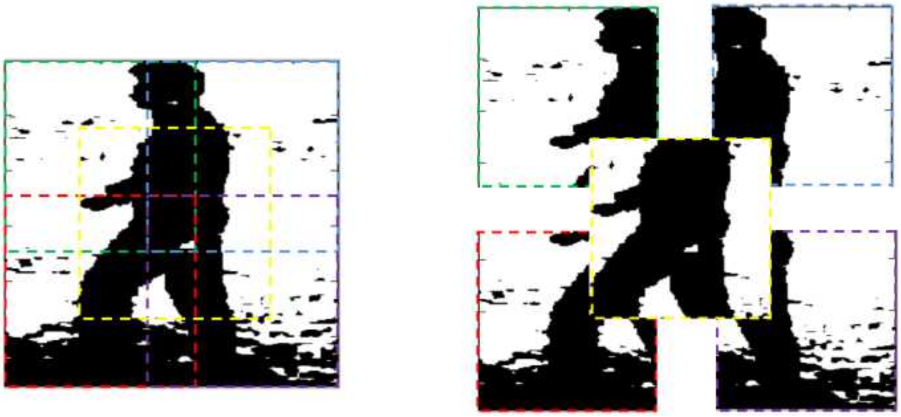

Fig. 4. Image cropping strategy.

original image.

5.2 The training process

Database preprocessing: in the time domain of convolutional neural network model learning, we need to obtain the optical flow image of video continuous frame in advance. Each optical flow image is composed of X vector image and Y vector image, and the optical flow image is shown in Fig. 4. The data layer sends 10 consecutive optical stream images at a time, that is, the image channel is 20. After random clipping, the data size is 224×224×20.

Data set: the purpose of training the network model is to extract time-domain features for fusion with C3D network model features. Therefore, the same data set as 3D CNN is adopted to divide UCF101 data set into a training data set and a test data set.

Pre-training model: when the training data set is small, pre-training has been proved to be effective in network initialization. Therefore, the network model obtained from ImageNet data set training is adopted for network initialization.

Network training: during the network training, the data size will change correspondingly according to the characteristics of different network layers. The edge-adding processing in the convolution layer will not change the size of the image, but the number of filters in each convolution layer will change, so the number of channels in the image will change with the number of filters. In the pooling layer, the purpose is to sample the image features, so the number of image channels will not change, and the size of image features will be halved.

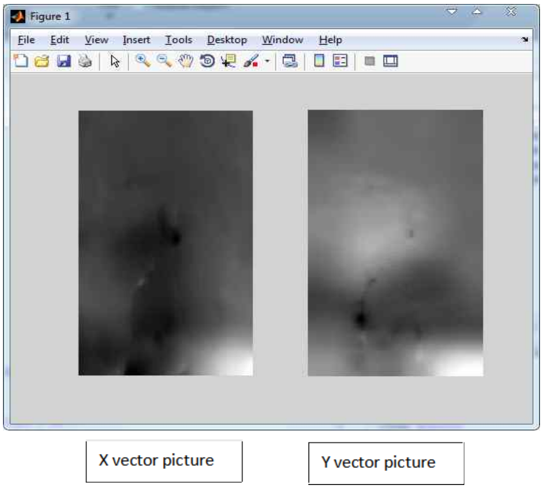

Fig. 5. Optical flow image.

5.3 Experimental result

In order to verify the effectiveness of this algorithm, a large number of comparative experiments were made on the published UCF101 database. Most of the samples of the UCF101 database are collected in movie clips, network resources or video surveillance. The resulting 101 behaviors are very close to the natural scene. And the experiment is run in MATLAB 2010b implementation.

The method of 3D CNN, Vgg16, feature fusion based on UCF101 database identification performance comparison shown in Table 2. The accuracy of feature fusion is greatly improved compared with 3D convolution network. The weight setting strategy has the highest accuracy of weighted feature fusion. The serial feature fusion has a slightly higher accuracy than the weighted feature fusion of the first weight setting strategy. However, because the feature dimension is doubled after serialization, the feature size is increased, so it takes more time to run.

Table 2. Four feature recognition results

3D CNN[15]	Dimension	Accuracy [%]
Two-stream CNN[15]	4096	57.4
P-CNN[16]	4096	58.0
Spatio-temporal CNN[17]	4096	65.4
Vgg16[18]	4096	79.2
$fusion_{concat}$	4096	68.6
$fusion_{weight1}$	8192	82.2
$fusion_{weight2}$	4096	82.4

6. CONCLUSION

Although there has been a lot of research and progress in video behavior recognition, due to camera motion, partial occlusion, complex background and large intra-class differences, video behavior recognition in real scenes still exists. More problems need to be solved and improved. Future research on behavior recognition methods can be carried out in the following aspects:

1) The 3D convolutional neural network based on motion trajectory now has a low trajectory extraction on the first 3D convolutional layer, resulting in a lower final recognition accuracy. In the future, it is hoped that improvements can be made on this basis. The 3D convolution layer is processed to improve accuracy.

2) Traditional artificial features are usually designed based on human prior knowledge and lack certain generalization capabilities. In addition, underlying artificial features are directly used for behavior recognition, and there are problems of insufficient semantic gap and discriminative ability. In recent years, although the deep learning-based behavior recognition method has made some progress, most methods still cannot fully learn the space-time characteristics in video. The advantages and disadvantages of the above two methods can be fully considered, and a new behavior recognition architecture is constructed.

3) At present, most of the behavior recognition methods based on deep learning methods ignore the inherent differences between video time domain and airspace. Therefore, designing a more effective deep learning network structure to better learn the space-time information in video behavior is also a research direction in the future.

REFERENCE

[1] R.T. Collins, A.J. Lipton, T. Kanade, H. Fujinyoshi, D. Duggins, Y. Tsin, et al., *A System for Video Surveillance and Monitoring*, Defense Advanced Research Projects Agency Image Understanding under Contract DAAB07-97-C-J031 and the Office of Naval Research under Grant, N00014-99-1-0646, 2000.

[2] Y. Zheng, Q.Q. Chen, and Y.J. Zhang, “Deep Learning and Its New Progress in Target and Behavior Recognition,” *Chinese Journal of Image and Graphics*, Vol. 19, No. 2, pp. 175-184, 2014.

[3] Q.J. Xu and Z.Y. Wu, “Research Progress on Behavior Recognition in Video Sequences,” *Journal of Electronic Measurement and Instrument*, Vol. 28, No. 4, pp. 343-351, 2014.

[4] Q. Lei, D.S. Chen, and S.Z. Li, “New Progress in Human Behavior Recognition in Complex Scenes,” *Computer Science*, Vol. 41, No. 12, pp. 1-7, 2014.

[5] P.P. Peng, *Image Classification Based on Set Representation,* Master’s Thesis of Harbin Engineering University, 2016.

[6] J. Ma, *Research and Implementation of Action Recognition Based on Pose and Skeleton*, ShanDong University of Control Engineering Language Institute, 2018.

[7] J. Arunnehru, G. Chamundeeswari, and S.P. Bharathi, “Human Motion Recognition Using 3D Convolutional Neural Network with 3D Motion Cuboids in Surveillance Videos,” *Proceeding of International Conference on Robotics and Smart Manufacturing*, pp. 471-477, 2018.

[8] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao, “Towards Good Practices for Very Deep Two-stream Conv Nets,” *Computer Science*, Vol. 10, No. 2, pp. 75-78, 2015.

[9] S. Ji, W. Xu, M. Yang, and K. Yu, “3D Convolutional Neural Networks for Human Action Recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, Issue 1, pp. 221-231, 1996.

[10] S. Ji, M. Yang, and K. Yu, “3D Convolutional

Neural Networks for Human Action Recognition," *Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221-231, 2013.

[11] G. Cheron, I. Laptev, and C. Schmid, "P-CNN: Posed-based CNN Features for Action Recognition," *Computer Vision*, Vol. 10, No. 10, pp. 3218-3226, 2015.

[12] M. Simonyan and A. Zisserman, "Two-steam Converlution Network for Action Recongnition in Videos," *Computational Linguistics*, Vol. 1, No. 4, pp. 568-576, 2014.

[13] N. Zhang and E.J. Lee, "Human Action Recognition Based on An Improved Combined Feature Representation," *Journal of Korea Multimedia Society*, Vol. 21, No. 12, pp. 1473-1480, 2018.

[14] P. Matikainen, M. Hebert, and S.R. Trajectons, "Action Recognition through the Motion Analysis of Tracked Features," *Proceeding of IEEE International Conference on Computer Vision Workshops*, pp. 514-521, 2009.

[15] K. Simonyan and A. Zisserman, "Two-stream Convolutional Networks for Action Recognition in Video," *Advances in Neural Information Processing Systems*, pp. 568-576, 2014.

[16] G. Chéro, I. Laptev, and C. Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," *Proceeding of the IEEE International Conference on Computer Vision*, pp. 3218-3226, 2015.

[17] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale Video Classification with Convolutional Neural Networks," *Proceeding of Conference on Computer Vision and Pattern Recognition*, pp. 1725-1732, 2014.

[18] T. Du, L. Bourdev, R. Fergus, and Y. Qiao, "Towards Good Practices for Very Deep Two-Stream Conv Nets," *Proceeding of IEEE International Conference on Computer Vision*, pp. 4489-4497, 2015.

Tingting Wu

received her B. s. at Dalian Polytechnic University in China (2011-2015). Currently, she is studying in Department of Information and Communication Engineering, Tongmyong University, Korea for M. S. Her main research areas are image processing and pattern recognition.

Eung-Joo Lee

received his B. s., M. s. and Ph.D. in Electronic Engineering form Kyungpook National University, Korea, in 1990, 1992 and Aug. 1996, respectively. Since 1997 he has been with the Department of Information & Communications Engineering, Tongmyong University, Korea, where he is currently a professor. From 2000 to July 2002, he was a president of Digital Net Bank Inc. From 2005 to July 2006, he was a visiting professor in the Department of Computer and Information Engineering, Dalian Polytechnic University, China. His main re-search interests include biometrics, image processing, and computer vision.