Journal of Korea Multimedia Society Vol. 23, No. 8, August 2020(pp. 977-985)
https://doi.org/10.9717/kmms.2020.23.8.977

Human Motion Recognition Based on Spatio-temporal Convolutional Neural Network

Zeyuan Hu[†], Sange-yun Park[††], Eung-Joo Lee[†††]

ABSTRACT

Aiming at the problem of complex feature extraction and low accuracy in human action recognition, this paper proposed a network structure combining batch normalization algorithm with GoogLeNet network model. Applying Batch Normalization idea in the field of image classification to action recognition field, it improved the algorithm by normalizing the network input training sample by mini-batch. For convolutional network, RGB image was the spatial input, and stacked optical flows was the temporal input. Then, it fused the spatio-temporal networks to get the final action recognition result. It trained and evaluated the architecture on the standard video actions benchmarks of UCF101 and HMDB51, which achieved the accuracy of 93.42% and 67.82%. The results show that the improved convolutional neural network has a significant improvement in improving the recognition rate and has obvious advantages in action recognition.

Key words: Convolutional Neural Network; Deep Learning; Human Activity Recognition; Spatio-temporal Convolutional Neural Network.

1. INTRODUCTION

Human action recognition is a hot research topic in both academic and industry research [1-4]. The technique is widely used in human-computer interaction (HCI)[5-7], human motion analysis[8,9], video surveillance[10,11], and content-based image storage and retrieval. Human action recognition aims to recognize and understand human motion from video stream and image sequences. Since the diversity of real motion scenarios, such as luminance, angles of cameras, and complex environment, human action recognition is still a challenging task. In different motion scenarios, the same action might exhibit different appearance. Even in a fixed motion scene, because of the large degree of freedom in the expression of human motion, the same movement has great visual differences in direction and angle. In addition, some external factors, such as occlusion, will also affect the performance of human motion recognition.

General human action recognition algorithms always extract feature descriptors, which are trained to learn a classifier (such as SVM) to achieve human action classification and recognition. In order to achieve high recognition accuracy, the extracted feature descriptors should be robust. However, feature descriptors of conventional algorithms were designed manually, which required high prior knowledge about trained and tested data sets. These algorithms were generally complicated for generalization. In addition, since feature descriptors designed manually are always task-driven, which are dependent on specific tasks. Thus, man-

※ Corresponding Author : Eung-Joo Lee, Address: 428, Sinseon-ro, Nam-gu, Busan, Korea, TEL : +82-51-629-1143, E-mail : ejlee@tu.ac.kr
Receipt date : Jun. 23, 2020, Approval date : Jul. 2, 2020

[†] Dept. of Information Communication Engineering, Tongmyong University E-mail : dlhzy410@126.com
[††] Academic Affairs and Register Team, Silla University E-mail : sypark@silla.ac.kr
[†††] Dept. of Information Communication Engineering, Tongmyong University

made feature descriptors cannot guarantee that essential features can be obtained from human motion sequences. Thus, designing robust feature descriptors of human motion is significant for action recognition. In recent years, deep learning-based algorithms have shown impressive performance in image processing, such as image classification[12, 13], image recognition, and image retrieval. These algorithms mimic human neural network, which consists of multiple layers to combine low-level features to form abstract high-level features. As a well-known representation of neural network, convolution neural network-based algorithms have demonstrated the impressive performance compared with traditional approaches. Through supervised and semi-supervised learning schemes, CNN based methods can learn the essential features of data, which outperform traditional methods by a large margin. Convolutional neural network has certain scale invariance and translation invariance. Local weight sharing network structures can reduce training parameters in image processing and has better performance. CNN-based algorithms also play an important role in human action recognition. The general pipeline of CNN-based human action recognition can be summarized as follows: Human action video stream or image sequences are fed into convolution neural network as training data. Multiple network layers will extract multi-scale features and fully-connected layers will integrate these features. Afterward, Softmax function will be used for human action classification and recognition.

2. RELATED WORK

Human action recognition is a hot research topic in both academic and industry domains [14-18]. Traditional algorithms mainly aimed to analyze the RGB image sequences. Shotton et al.[21] utilized Harris detector and Gabor detector to detect spatial interest points, which will be constructed for 3D gradient histogram (HOG3D) to represent features. The authors proposed a human motion recognition method based on color spatiotemporal interest points. Kovashka et al.[19] proposed a human action recognition based on learning discriminative space-time neighborhood features. Instead of learning the predefined local descriptors, the proposed method aimed to learn the shape information of space-time neighborhoods to discover the most discriminative cues for human action recognition. A class-specific distance function was introduced to recognize the most informative features to classify different human actions. Hussein et al.[20] proposed a temporal hierarchy of covariance descriptors to describe human action. The proposed method leverage 3D skeletion sequences extracted by depth camera, such as Kinect. In order to leverage the information between joint movement and time, the authors leverage covariance matrices of some sequences from a designed hierarchical scheme. Skeleton joint point recognition is also an effective algorithm for action recognition. Comprehensive experiments demonstrate that most action can be recognized by skeleton joint point recognition. Ellis et al. proposed to train a classifier based on logistic regression by using the delayed perception learning algorithm. The classifier can automatically determine the key posture from the 3D joint position sequence, which reduces the time of action recognition.

Deep learning-based algorithms aimed to construct an effect network recognition framework. Simonyan et al.[22] proposed a two-stream network structure, which proved that the convolutional neural network trained by inter-frame optical flow characteristics can still achieve good performance under the condition of limited data sets. He et al.[23] leverage spatial pyramid pooling to extract features from the last convolution layer. Wang et al.[24] constructed a maximum pooling network of three-dimensional convolution kernel to automatically recognize RGB-D video stream.

In[25], Wang et al. presented very deep two-stream convolution neural network based on the several classical neural networks. In order to solve the problem that 2D convolutional neural network cannot handle video sequences, Ji et al.[26] proposed 3D convolutional neural network based on 2D network. Tang et al.[27] proposed C3D network based on VGG. The experimental results shown that the network can achieve the best performance when setting the convolution kernel as $3\times3\times3$.

This paper proposes a network architecture that combines batch zeroing transformation and Goog LeNet network model and applies it to the field of video human action recognition. Compared with the traditional deep convolutional neural network, the training algorithm and network structure are improved in two aspects.

3. PROPOSED METHOD

In this section, we propose a novel spatio-temporal network for human action recognition, where batch normalization and GoogLeNet are combined to achieve this task.

3.1 Optical flow feature

Optical flow aims to recognize motion based on the spatial-temporal transformation and correlation of gray levels in image sequences. In general, let $I(x,y,t)$ denotes video sequence, where $x(t)$, $y(t)$ denotes the (x,y) pixel of the time. Based on the conservation of brightness, we can obtain the following formula.

$$\frac{d}{dt}I(x(t),y(t),t)=0 \tag{1}$$

For the trajectory points of each pixel on a sequence image, the instantaneous velocity vector field $U(x,y)=(u_1(x,y),u_2(x,y))$ denotes the optical flow. Then, the vector field satisfies the optical flow constraint equation:

$$I_1(x,U)-I_0(x)=0 \tag{2}$$

According to the gray level conservation and global smoothing constraints, the calculation of optical flow field is transformed into the minimization of capacity function.

$$E_{(u)}=\int\left(\nabla I\cdot U+\frac{\delta}{\delta_t}I\right)^2+\alpha\left(|\nabla u_1|^2+|\nabla u_2|^2\right) \tag{3}$$

where $|\nabla u_1|^2+|\nabla u_2|^2$ denotes the smooth constraint function. α is a weight factor that affects gray level conservation and global smoothing. The energy function is expressed by the data term of L_1 norm and the regular term of the total variation of optical flow, and taking into account the convex relaxation minimization function, $E_{(u)}$ can be reorganized as:

$$E_\theta(U,V)=\int|\nabla U_1|+|\nabla U_2|+\frac{1}{2\theta}|U-V|^2+\lambda|\rho(V)| \tag{4}$$

When θ is a small constants with smaller value, the minimization formula (4) is equivalent to the minimization formula (3), which minimizes the energy function by fixing U or V at one time and optimizing another variable. Optical flow estimation can be obtained by repeating the process. The example samples from HMDB51 data set is shown in Fig. 1.

3.2 Batch normalization

Assuming that denotes the input of network and b denotes the bias. The input set of the training images is denoted as $X=\{x_1,x_2,\cdots,x_N\}$, then the mean value of the training set is $E[X]=\frac{1}{N}\sum_{i=1}^{N}X_i$ Since the updating of offset b will be cancelled in

Fig. 1. The optical flow of the movement "sit" from HMDB51 data set.

normalization, that is, the updating of b and the combined effect of normalization operation do not change the input of the network layer, and then the loss function of the network layer will not be changed. As the whole network training iteration proceeds, b will increase continuously, but the loss function will remain unchanged. By normalizing the input of the network layer, we hope to weaken the problem of internal covariate migration, which may bring new problems. Therefore, when calculating the gradient of loss function to model parameters, it is necessary to consider the dependence of normalization operation on network parameters.

Let x denotes the input vector of network, X denotes the training set. We perform independent batch normalization instead of joint normalization for each dimension of input.

$$\hat{x}^{(k)} = \frac{x^{(k)} - E[x^{(k)}]}{\sqrt{Var[x^{(k)}]}} \tag{5}$$

Where $x^{(k)}$ denotes the k-th dimension of the input sample dataset. $E[x^{(k)}]$, $Var[x^{(k)}]$ represents the expectation and variance of the input, respectively. In random gradient training, micro-batch samples are used to train, and each layer is calculated on each micro-batch sample to estimate the mean and variance of the layer. Therefore, the neural network statistics calculated in batch normalization processing can be used in gradient back propagation.

3.3 Spatio-temporal convolutional neural network

The video stream consists of two components: time and space. Each individual frame contains the appearance information of the scene and object. In the time part, the motion information of the camera and the object is included. The model of spatio-temporal dual-flow network is shown in the Fig. 2. RGB image is input to the spatio-temporal network, and time flow field is input to the spatio-temporal dual-flow network.

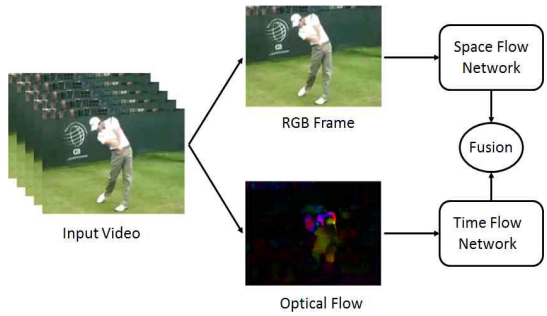


Fig. 2. The architecture of deep spatio-temporal network.

We leverage GoogLeNet as our basic network architecture. In this paper, we leverage the inception model structure with convolution sliding window filter sizes of 1×1 and 3×3. In the inception model structure, the output of convolution filters with different branches is fed into a filter cascade layer, and then all the outputs are expanded into a column vector as the input of the next layer. Compared with the original model, the dimension-reducing Inception model structure adds a convolution filter operation with sliding window size of 1×1. By setting the total number of convolution output feature maps of these 1×1 convolution filters to a smaller number, the dimension-reducing effect is achieved, and then the computational complexity is reduced in the 3×3 convolution filter with larger access overhead.

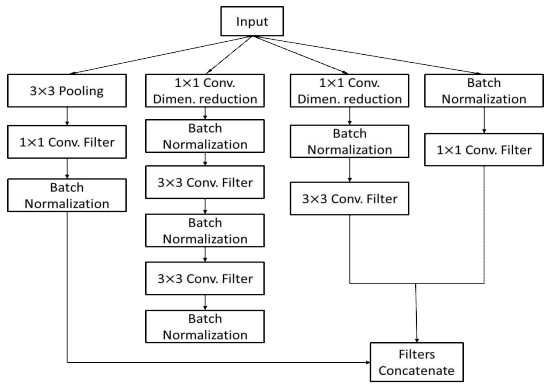


Fig. 3. The Inception model with batch normalization processing.

After the pooling filter layer, a 1×1 convolution filter is added to reduce the total number of the branch filters, thus reducing the dimension of the output feature graph. The basic architecture of Inception model with batch normalization processing is shown in Fig. 3. The whole architecture consists of 3 convolution blocks, 3 maximum pooling blocks and 10 Inception blocks. Finally, human action classification and recognition can be achieved by Softmax function.

4. EXPERIMENT AND ANALYSIS

4.1 Model Training

Available data sets for human action recognition, such as HMDB51, UCF101, are limited by their scale compared with other well-known dataset such as ImageNet. In order to alleviate over fitting, we first conduct data enhancement. Since the random clipping technology tends to select the central block of the image, it is easy to cause over-fitting. Therefore, we cut the edge and center blocks of the image frame to enhance the scale diversity. Specifically, we fix the input data to 256×340, then randomly select a candidate cutting size from the set$\{256,224,192\}$ for clipping, and finally adjust the clipped blocks to 224×224.

In pretraining, the deep convolution neural network is pre-trained on the ImageNet data set. Since the input of space network is RGB image, the network can be initialized directly by using the ImageNet pre-training model, and the optical flow field of 10 frames stacked at the input of time network, so some network adjustments are needed. In our implementation, we first use OpenCV to extract the optical flow field of action video, and then use linear transformation to discretize the optical flow to $[0,255]$ interval. Finally, the filter of the first layer of the spatial network model trained by ImageNet is averaged in the channel, and the average structure is duplicated 10 times as the initialization of the time network.

Our experiment is conducted on Caffe framework. In our implement, for space network, the initial learning rate is 0.001, and the learning rate will be 0.0001 after 1800 iterations. For time network, the initial learning rate is 0.003, and the learning rate will be 0.0003 after 15000 iterations. We set the weight decay and momentum to 0.0002 and 0.9, respectively.

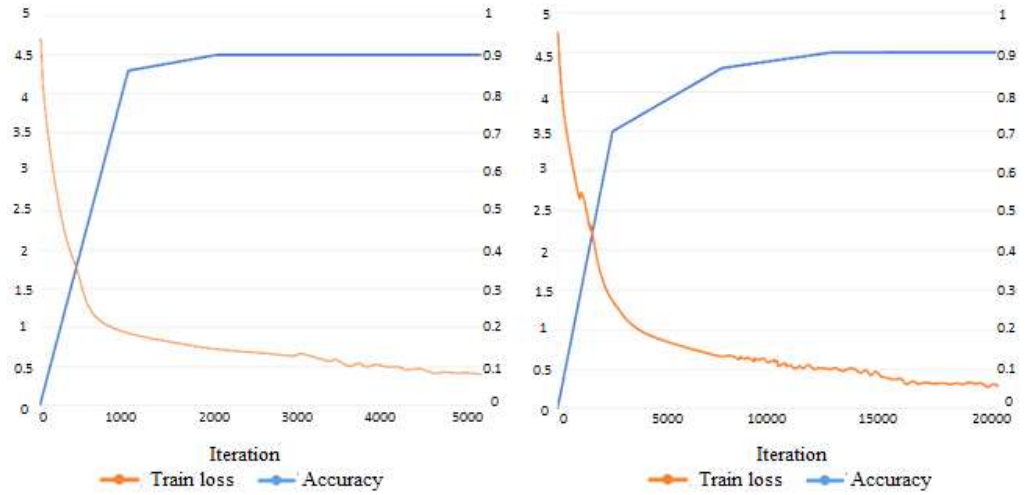
4.2 Comparative Study

We leverage UCF101 data set in our work. Dropout layer is added in our neural network architecture to avoid overfitting. We explore the influence of dropout_ratio of Dropout layer on the recognition accuracy in the constructed spatio-temporal motion recognition network. The result is shown in Table 1. As we can see from Table 1, when the dropout rate of time network is 0.7, the recognition rate is 0.17% and 0.35% higher than that of 0.4 and 0.6. When the dropout rate of space network is 0.8, the recognition rate is 1.03% and 0.45% higher than that of 0.4 and 0.6, respectively.

Fig. 4(a) and Fig. 4(b) show the training convergence of spatiotemporal networks. We can see form Fig. 4(a), on the spatial network, when the number of training iterations reaches 1000, the accuracy is close to 86%, and the loss value of training decreases rapidly. When the number of training times reaches 2000, the accuracy remains above 90% and the loss value is below 0.1. With the iteration, the convergence tends to be stable. Similarly, on the time network, when the number of training

Table 1. The effect of different parameters on recognition accuracy

Network	(dropout ratio) accuracy	(dropout ratio) accuracy	(dropout ratio) accuracy
Time Network	(0.4) 86.52%	(0.6) 86.34%	(0.7) 86.69%
Space Network	(0.4) 82.54%	(0.6) 83.12%	(0.8) 83.57%



(a) Space network training iteration (b) Time network training iteration

Fig. 4. The training convergence of spatio-temporal networks.

iterations reaches 7500, the accuracy is close to 85%, and the loss value of training decreases rapidly. When the number of training times reaches 12500, the accuracy remains above 90% and the loss value is below 0.08.

we use the standard training splits and protocols provided as the original evaluation scheme. For UCF101, we found that, In the UCF-101 motion database, the recognition rate of improved three-dimensional convolutional neural network was 89.4%, And by shifting the structural framework of learning, it's found that it saves a lot of time compared to traditional methods. The experiment result showed in table I, and we do some comparison with some other method, the result shows that The 3D Hubrid Model algorithm, as the best behavior recognition algorithm before the depth of learning, has excellent results and good robustness.

It can be seen from the experimental results that the improved three-dimensional convolutional neural network deep learning method has higher recognition efficiency than the k-proximity and support vector machine methods in databases, and has better learning ability under the same training samples.

In our implement, we leverage linear weighting method to fuse the classification results of spatio-temporal networks, and the recognition structure is shown in the Table 2. Each experiment is conducted for 10 times and we calculate the average value as the final recognition accuracy. We use linear weighted fusion to get the final recognition rate. When the weight of recognition confidence of spatial network and temporal network classification is set to 1:1, the performance of the fusion space-time convolution neural network is better than other cases. In human motion recognition tasks, the fusion of spatiotemporal two-stream convolutional neural networks can effectively improve the accuracy of individual networks in recognition.

Further, we conduct comparative experiments

Table 2. The recognition accuracy of the spatio-temporal network

Database	UCF101			HMDB51		
Space,Time	1:1	1:1.2	1:1.5	1:1	1:1.2	1:.5
Accuracy	0.9342	0.9316	0.9276	0.6853	0.6798	0.6715

Table 3. The recognition accuracy of the spatio-temporal network

Methods	UCF101	HMDB51
Improved dense	0.8587	0.5721
IDT with higher-dimensional	0.8793	0.6116
Two-stream [25]	0.8846	0.5932
Very deep two-stream [28]	0.9133	0.6044
KVMF [29]	0.9310	0.6329
Proposed Method	0.9342	0.6783

with several well-known human action recognition algorithms, as shown in Table 3. Improved dense method uses a dense trajectory algorithm. IDT with higher-dimensional method is an enhanced version of traditional BOVW visual-word-bag model, which achieves higher dimensional feature coding. Two-stream achieves human action recognition based on constructing a two-stream spatio-temporal network model. However, the network is shallow. Further, very deep two-stream method is proposed to solve the shortcoming of original two-stream method. KVMF method uses video segments to intercept multiple 3D volumes as the input of the network, and uses the predictive vectors from each volume to represent the action categories it belongs to.

5. CONCLUSION

In this paper, we construct a network architecture combining batch normalization transformation with GoogLeNet model and apply it to human motion recognition from video stream. The improved network architecture is utilized to construct a spatio-temporal convolution neural network model to realize motion recognition. Spatial flow network obtains the appearance information of motion through RGB images of video stream, while temporal flow captures the motion information through optical flow fields between consecutive frames. Finally, the spatio-temporal network fuse both the appearance and motion information.

REFERENCE

[1] L. Xia, C.C. Chen, and J.K. Aggarwal, "View Invariant Human Action Recognition Using Histograms of 3D Joints," *Proceeding of Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pp. 20-27, 2012.

[2] A. Kovashka and K. Grauman, "Learning a Hierarchy of Discriminative Space-time Neighborhood Features for Human Action Recognition," *Proceeding of Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2046-2053, 2010.

[3] M.E. Hussein, M. Torki, M.A. Gowayyed, and M.E. Saban, "Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations," *Proceeding of International Joint Conference on Artificial Intelligence*, pp. 2466-2472, 2013.

[4] M.Y. Liu, H. Liu, and C. Chen, "Enhanced Skeleton Visualization for View Invariant Human Action Recognition," *Journal of Pattern Recognition*, Vol. 68, No. 1, pp. 346-362, 2017.

[5] Y.L. Song, D. Demirdjian, and R. Davis, "Continuous Body and Hand Gesture Recognition for Natural Human-computer Interaction," *Journal of ACM Transactions on Interactive Intelligent Systems*, Vol. 2, No. 1, pp. 4212- 4216, 2012.

[6] T. Santini, W. Fuhl, and E. Kasneci, "CalibMe: Fast and Unsupervised Eye Tracker Calibration for Gaze-based Pervasive Human-computer Interaction," *Proceeding of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 2594-2605, 2017.

[7] Y.K. Meena, H. Cecotti, K.W. Lin, A. Dutta, and G. Prasad, "Toward Optimization of Gaze-controlled Human-computer Interaction: Application to Hindi Virtual Keyboard for Stroke Patients," *Journal of IEEE Transactions on Neural Systems and Rehabilitation Engineer-*

ing, Vol. 26, No. 4, pp. 911-922, 2018.

[8] L.L. Chen, H. Wei, and J. Ferryman, "A Survey of Human Motion Analysis Using Depth Imagery," *Journal of Pattern Recognition Letters*, Vol. 34, No. 15, pp. 1995-2006, 2013.

[9] G.P. Dominguez, B. Taati, and A. Mihailidis, "3D Human Motion Analysis to Detect Abnormal Events on Stairs," *Proceeding of 2012 Second International Conference on 3D Imaging, Modeling, Processing, Visualization, and Transmission*, pp. 97-103, 2012.

[10] C.B. Jin, S.Z. Li, T.D. Do, and H. Kim, "Real- Time Human Action Recognition Using CNN over Temporal Images for Static Video Surveillance Cameras," *Proceeding of Pacific Rim Conference on Multimedia of Advances in Multimedia Information Precess*, pp. 330-339, 2015.

[11] B.Y. Wang, Y.L. Hu, J.B. Gao, Y.F. Sun, and B.C. Yin, "Laplacian LRR on Product Grassmann Manifolds for Human Activity Clustering in Multicamera Video Surveillance," *Journal of IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 27, No. 3, pp. 554-566, 2017.

[12] L. Hou, D. Samaras, T.M. Kurc, Y. Gao, J.E. Davis, and J.H. Saltz, "Patch-based Convolutional Neural Network for Whole Slide Tissue Image Classification," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2424-2433, 2016.

[13] S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk, and D. Stefanovic, "Deep Neural Networks Based Recognition of Plant Diseases by Leaf Image Classification," *Journal of Computational Intelligence and Neuroscience*, Vol. 2016, No. 6, pp. 1-11, 2016.

[14] S.J. Song, C.L. Lan, J.L. Xing, W.J. Zeng, and J.Y. Liu, "An End-to-end Spatio-temporal Attention Model for Human Action Recognition from Skeleton Data," *Proceeding of American Association for Artificial Intelligence Conference on Artificial Intelligence*, pp. 4263-4270, 2017.

[15] A. Iosifidis, A. Tefas, and I. Pitas, "Human Action Recognition Based on Multi-view Regularized Extreme Learning Machine," *Journal of Artificial Intelligence Tools*, Vol. 24, No. 5, pp. 1-11, 2015.

[16] J. Wang, Z.C. Liu, Y. Wu, and J.S. Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," *Proceeding of 2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1290-1297, 2012.

[17] 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.

[18] Z.Y. Hu, S.Y. Park and E.J Lee, "Human Action Recognition Based on Convolutional Neural Network," *Proceeding of 2018 Conference on Korea Multimedia Society*, pp. 233-235, 2018.

[19] A. Kovashka and K. Grauman, "Learning a Hierarchy of Discriminative Space-time Neighborhood Features for Human Action Recognition," *Proceeding of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2046-2053, 2010.

[20] M.E. Hussein, M. Torki, M.A. Gowayyed, and M.E. Saban, "Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations," *Proceeding of the Twenty-third International Joint Conference on Artificial Intelligence*, pp. 2466-2472, 2013.

[21] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, and A. Blake, "Real-time Human Pose Recognition in Parts from Single Depth Images," *Journal of Communications of the ACM*, Vol. 56, No. 1, pp. 116-124, 2013.

[22] K. Simonyan and A. Zisserman, "Two-stream Convolutional Networks for Action Recognition in Videos," *Proceeding of the 27th Inter-*

national Conference on Neural Information Processing Systems, pp. 568-576, 2014.

[23] K. He, X.Y. Zhang, S.Q. Ren, and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 37, No. 9, pp. 1904-1916, 2015.

[24] K. Wang, X.L. Wang, L. Lin, M. Wang, and W. Zuo, "3D Human Activity Recognition with Reconfigurable Convolutional Neural Networks," *Proceeding of the 22nd ACM International Conference on Multimedia*, pp. 97-106, 2014.

[25] L.M. Wang, Y.J. Xiong, Z. Wang, and Y. Qiao, "Towards Good Practices for Very Deep Two-stream ConvNets," *Journal of Computer Science-Computer Vision and Pattern Recognition*, Vol. abs/1507.02159, pp. 1-5, 2015.

[26] S.W. Ji, W. Xu, M. Yang, and K. Yu, "3D Convolutional Neural Networks for Human Action Recognition," *Journal of IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 35, No. 1, pp. 221-231, 2013.

[27] T. Du, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," *Proceeding of the IEEE International Conference on Computer Vision*, pp. 4489-4497, 2015.

[28] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *Journal of Computer Science,* Vol. abs/1409.1556, pp. 1-14, 2015.

[29] W. Zhu, J. Hu, G. Sun, X. Cao, and Y. Qiao, "A Key Volume Mining Deep Framework for Action Recognition," *Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1991-1999, 2016.

Zeyuan Hu

was born in Dalian, Liaoning. P.R. China, in 1992. He received the bachelor's degree in Automation from Qingdao Institute of Technology, P.R. China(2011-2015). He received the master's degree in Information Communication Engineering from Tongmyong University, Busan, Korea(2016-2018). Currently, he has been studying for his doctoral degree in the Department of Information and Communications Engineering in Tongmyong University, Korea. And he is majoring in image processing and pattern recognition.

Sang-yun Park

received his B.S(1992-1995), in Physics, M,S(1996-1998), in multimedia Application from Kyungsung University, Korea, in 1996, Aug. 1998, respectively, and Ph. D.(2006-2011) in Information & Communication Engineering from Tongmyong University, Korea. in Aug. 2011. His main research interests include computer vision, neural network, pattern recognition. image processing, HCI(human computer interaction) and biometrics.

Eung-Joo Lee

received his B. S., M. S. and Ph. D. in Electronic Engineering from Kyungpook National University, Korea, in 1990, 1992, and Aug. 1996, respectively. Since 1997 he has worked 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 research interests include biometrics, image processing, and computer vision.