JANT 한국항행학회논문지
Journal of Advanced Navigation Technology

응용 및 융합 기술

J. Adv. Navig. Technol. 24(2): 155-162, Apr. 2020

스켈레톤 조인트 매핑을 이용한 딥 러닝 기반 행동 인식

Deep Learning-based Action Recognition using Skeleton Joints Mapping

타스님 · 백중환*
한국항공대학교 항공전자정보공학부

Nusrat Tasnim · Joong-Hwan Baek*
School of Electronics and Information Engineering, Korea Aerospace University, Gyeonggi-do, 10540, Korea

[요 약]

최근 컴퓨터 비전과 딥러닝 기술의 발전으로 비디오 분석, 영상 감시, 인터렉티브 멀티미디어 및 인간 기계 상호작용 응용을 위해 인간 행동 인식에 관한 연구가 활발히 진행되고 있다. 많은 연구자에 의해 RGB 영상, 깊이 영상, 스켈레톤 및 관성 데이터를 사용하여 인간 행동 인식 및 분류를 위해 다양한 기술이 도입되었다. 그러나 스켈레톤 기반 행동 인식은 여전히 인간 기계 상호작용 분야에서 도전적인 연구 주제이다. 본 논문에서는 동적 이미지라 불리는 시공간 이미지를 생성하기 위해 동작의 종단간 스켈레톤 조인트 매핑 기법을 제안한다. 행동 클래스 간의 분류를 수행하기 위해 효율적인 심층 컨볼루션 신경망이 고안된다. 제안된 기법의 성능을 평가하기 위해 공개적으로 액세스 가능한 UTD-MHAD 스켈레톤 데이터 세트를 사용하였다. 실험 결과 제안된 시스템이 97.45 %의 높은 정확도로 기존 방법보다 성능이 우수함을 보였다.

[Abstract]

Recently, with the development of computer vision and deep learning technology, research on human action recognition has been actively conducted for video analysis, video surveillance, interactive multimedia, and human machine interaction applications. Diverse techniques have been introduced for human action understanding and classification by many researchers using RGB image, depth image, skeleton and inertial data. However, skeleton-based action discrimination is still a challenging research topic for human machine-interaction. In this paper, we propose an end-to-end skeleton joints mapping of action for generating spatio-temporal image so-called dynamic image. Then, an efficient deep convolution neural network is devised to perform the classification among the action classes. We use publicly accessible UTD-MHAD skeleton dataset for evaluating the performance of the proposed method. As a result of the experiment, the proposed system shows better performance than the existing methods with high accuracy of 97.45%.

Key words : Action recognition, Deep learning, CNN, End-to-end skeleton joints mapping.

https://doi.org/10.12673/jant.2020.24.2.155

Received 24 March 2020; **Revised** 25 March 2020
Accepted (Publication) 16 April 2020 (30 April 2020)

***Corresponding Author; Joong-Hwan Baek**

Tel: +82-2-2209-3671
E-mail: biscopro@naver.com

Ⅰ. Introduction

Digital devices such as computers, smart phones, cameras are now becoming an essential part of our daily life. The main motive of our research is to provide easy and comfortable methods for interacting with those machines. With the improvements of the research, the form of interaction with those devices has also updated. In previous times, the most common form of communication devices was keyboard and mouse. Now, we are expecting more reliable and pleasant ways to control our machines including computer vision-based face, iris, voice, gesture or action recognition. Action recognition is considered one of the most demanding ideas for contacting with the devices. It is getting more popular among the researchers due to its remarkable attributions in numerous fields for instance computer vision, image processing, and pattern recognition. The invention of low cost, easy to use and portable sensors along with some efficient data capturing tools provides different modalities of action detection as well as classification dataset (RGB, depth, skeleton and inertial) that are commonly used these days. A wide range of applications like gesture recognition, smart surveillance systems, home monitoring, identity recognition, game control, robotics, and ease human-machine interaction is spreading rapidly [1],[2]. Gesture or action identification plays more importance in this extends. A gesture is a form of non-verbal communication that is done by hand, fingers, arm or other parts of the human body. The hand gesture is the most popular in the areas of gesture recognition that can be divided into two board groups [3]; static and dynamic hand gestures. The static hand gesture mainly focuses on the information of a single image whereas the spatial-temporal feature is the major properties of the dynamic hand gesture as shown in Fig. 1.

In this paper, we design an algorithm for discriminating various types of human actions performed by different parts of the human body. Initially, we generate dynamic images using for all actions by mapping between different joints information of the neighboring frames. The spatio-temporal images along with three different views are fed into the networks for extracting meaningful features and then fused them to improve the classification rates. A modified version of the AlexNet is introduced for the purpose of the classification among the 27 action classes.

In section II, we try to illustrate some state-of-arts techniques related to action recognition and skeleton-based action detection and classification. In section III, we describe our proposed methodology along with end-to-end skeleton joints mapping, data augmentation, and modification of deep convolution neural network. Experimental results are shown in section IV and finally we include the conclusion in section V.

(a) Static Gestures (b) Dynamic Gesture

그림 1. 손 제스처
Fig. 1. Hand Gestures.

Ⅱ. Related Works on Action Recognition

An action is a process performed by a group of motions or frames that represent what a person is doing for instance walking, waving, clapping, etc. Normally, the response of performing an action lasts no more than a few seconds. Modern technology gifts us a variety of sensors (RGB camera, depth camera, RealSense, Microsoft Kinect Sensor) that provide accurate datasets for RGB, depth, inertial and skeleton modalities. By using those data, researchers are continuing their works for building effective and efficient algorithms using various learning approaches mainly machine learning and deep learning. Dollar et al. [4] illustrated an efficient method using a temporal Gabor filter and a spatial Gaussian filter for detecting spatio-temporal interest points (STIPs). Then, the authors proposed some other STIP detectors and descriptors for improving the results. Wu et al. [5] suggested a method by combining both local and global feature representations for action recognition. They used temporal local feature descriptor and motion descriptor named bag of corrected poses (BoCP) and extended motion history image (extended-MHI) respectively for their classification. Ahmed et al. [6] explained some features like body silhouette feature, optical flow feature and combined feature and then used hidden markov model (HMM) for action identification. Xia et al. [7] generated histograms of 3D joint locations for recognizing human action and HMM for classification. Luo et al. [8] discussed a method where the temporal pyramid matching approach (ScTPM) was used for feature representation and support vector machine (SVM) was used for classification. Megavannan et al. [9] proposed a feature extraction method for depth images named Hu Moments and a silhouette bonding box and used SVM for

classification. For action recognition, Trelinski et al. [10] proposed a convolution neural network (CNN) model where onto an orthogonal Cartesian plane consecutive depth maps and depth maps are projected. In [11], Wang et al. described a new method from depth maps using weighted hierarchical depth motion maps (WHDMM) and three-channel deep convolutional neural networks (3ConvNets) for human action recognition. Simonyan et al. [12] introduced two-stream convolutional neural networks for action recognition. They improved the model in three separate ways. At first, they proposed a two-stream ConvNet architecture by combining spatial and temporal networks. Secondly, ConvNet was trained on multi-frame dense optical flow and finally applied this method to different action classification datasets.

Many researchers have spent their valuable time in developing algorithms for action recognition based on the skeleton joints information in the 3-dimensional coordinate system of the human body. In the early days, many handcrafted methods were proposed for the extraction of distinctive features from the skeleton data of various actions in order to perform the recognition. Feature extraction for Depth and RGB sequences needs more computation than skeleton data in terms of time and computing resources. Most of the existing models used skeleton joints information represented in the spatial domain based on handcrafted features. Several methods used HMM for capturing the temporal information from the skeleton data in the early days. After that, deep learning networks like recurrent neural network (RNN) or CNN were used largely for skeleton-based action recognition in last the few years.

In [13], Li et al. proposed a skeleton transformation module to select skeleton joints automatically and designed a CNN with 7-layers for action classification. In [14], Hussein et al. introduced a covariance descriptor to encode the relation between joint movement and time and then used SVM for classification. Du et al. [15] represented the skeleton joints into a matrix form and then converted it into images for the CNN network as input. Wang et al. [16] introduced a method named Joints Trajectory Map (JTM) where 3D skeleton sequences transformed into 2-dimensional images and CNN was used for classification. Hou et al. [17] illustrated an efficient method for action recognition with skeleton data named Skeleton Optical Spectra (SOS) where skeleton sequences which contain spatio-temporal information into color texture images and trained CNN for action classification. Li et al. [18] suggested a method named Joint Distance Map (JDM) which consists of a sequence frame to capture temporal information with different colors. In [16]-[18], Wang et al., Hou et al., and Li et al. represented the temporal information using HSB (Hue, Saturation and Brightness) color maps. For more illustration, the details representation is given in Fig. 2. In [19], J. Imran et al. suggested data augmentation for 3D skeleton joints information named 3-dimensional transformation and for classification purposes, RNN was designed named Bidirectional gated recurrent unit (BiGRU).

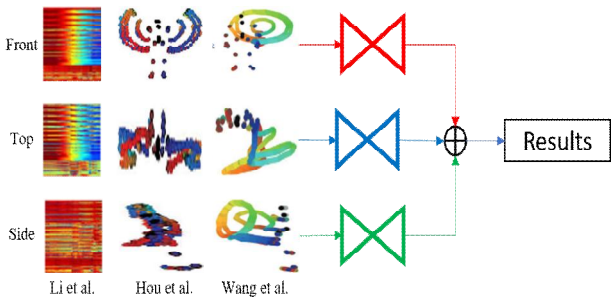


그림 2. 최신 행동 인식 시스템
Fig. 2. State-of-arts action recognition systems.

III. Action Recognition using Skeleton Joints Mapping

In this paper, we introduce a new method for action recognition using skeleton joints mapping information of 3-dimensional coordinate systems. First, we convert all the joints along *XY*, *YZ*, and *ZX*-axes of every frame in a video into a single frame by joining the line between the joints in neighboring frames. Then, updated AlexNet is used for discrimination among the action classes. In our system, we replace the mapping of skeleton joints using the HSB color model in [16]-[18] with the connecting lines between adjacent joints of frames using the RGB color model. The overall architecture of the proposed system is shown in Fig. 3.

The inputs to the system are the spatio-temporal images along with three different views (*XY*, *YZ*, and *ZX*-axes) representing with red, green and blue as shown in Fig. 3. These inputs are passed through the several layers of the networks for extracting

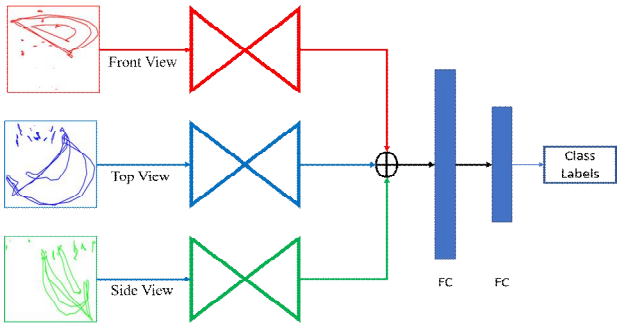


그림 3. 제안한 기법의 구조
Fig. 3. Architecture of proposed method.

meaningful features and then concatenated the outcomes from the front, top, and side views. Finally, two fully connected layers following with softmax layer are used for deciding the class labels of the performed actions. The proposed system consists of three basic parts. (1) end-to-end skeleton joints mapping, (2) data augmentation, and (3) network modification. The next three subsections explain the details representation of the end-to-end joints of all the frames in an action, data augmentation techniques, and network configurations.

We use UTD Multimodal Human Action Dataset skeleton dataset (UTD-MHAD) [20]. In this dataset, the coordinates of 20 different joints from the human body are extracted along XYZ-axes using Microsoft Kinect camera. Twenty different skeleton joints with their corresponding names are depicted in Fig. 4.

3-1 End-to-End Skeleton Joints Mapping

While an action is performed, the positions of the joints change in different views (along *XY*, *YZ*, and *ZX*-axes) from frame to frame in temporal direction over time. We map each joint between the neighboring frames by adding a line. Let us consider two frames F_{ij} and $F_{(i+1)j}$ where i is the index of two consecutive frames and j is the number of joints in each frame. Then the equation for joining frame F_{ij} and $F_{(i+1)j}$ in three different views (side, top, and front) is given in equation (1) - (3).

For front view (along *XY*-axes), the joints map representation can be expressed as

$$y - y_{i,j} = \frac{y_{(i+1),j} - y_{i,j}}{x_{(i+1),j} - x_{i,j}} \times x - x_{i,j} \tag{1}$$

where $i = 1, \cdots, n$; n is the length of an action and $j = 1, \cdots, m$; m is the number of joints.

Similarly, for side (along *YZ*-axes) and top (along *ZX*-axes) views, the equations can be written as

$$z - z_{i,j} = \frac{z_{(i+1),j} - z_{i,j}}{y_{(i+1),j} - y_{i,j}} \times y - y_{i,j} \tag{2}$$

and

$$x - x_{i,j} = \frac{x_{(i+1),j} - x_{i,j}}{z_{(i+1),j} - z_{i,j}} \times z - z_{i,j} \tag{3}$$

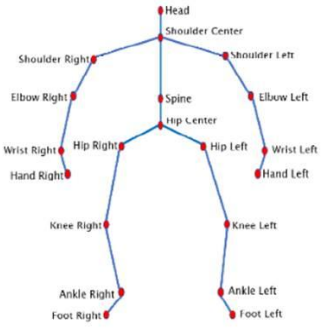


그림 4. 20개의 스켈레톤 조인트
Fig. 4. Twenty skeleton joints.

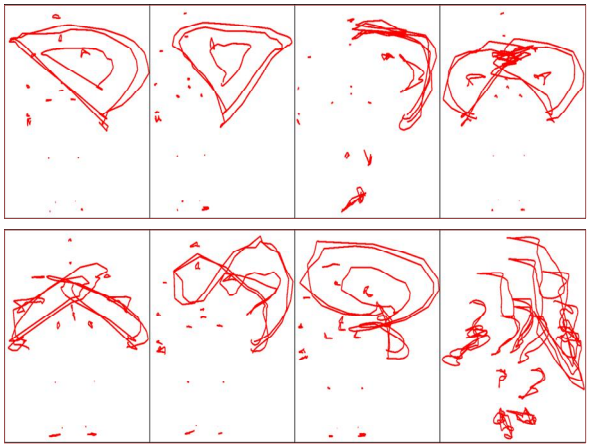
The main purposed of the spatial representation is to discriminate the actions using the texture features that changes from frame to frame. In order to explain the facilities of spatio-temporal representation, the mapping of eight different actions (swipe left, swipe right, wave, clap, arm cross, draw cross, draw circle clockwise, bowling) along *XY*, *YZ*, and *ZX*-axes are depicted sequentially in Fig. 5.

As shown in Fig. 5, some of the joints change along the temporal direction and some joints remain static as the actions are performed over time. The first three action classes swipe left, swipe right, and wave are performed using joints in the left hand in which mostly the joints named shoulder left, elbow left, wrist left, and hand left change from frame to frame. The rest of the joints remain static or a little change happened. The same things can be noticed for actions draw cross and draw circle clockwise that are done by using the joints of left hand. Some of the actions require both left hand and right hand joints. For actions clapping and arm cross needed both hands that make the movement of shoulder left, elbow left, wrist left, hand left, shoulder right, elbow right, wrist right, and hand right joints. Finally, the last action class (bowling) can be accomplished by using all the joints of the human body in which most of the joints change a lot.

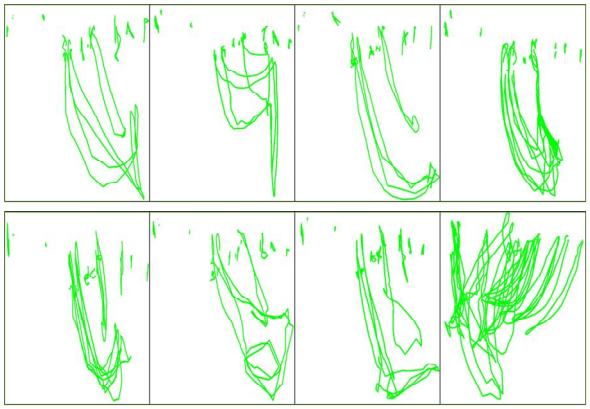
3-2 Data Augmentation

Due to the less amount of data, we consider the most popular type of data augmentation method for training the proposed DCNN models effectively. Since our main focus is to represent the joints into spatial maps that are dependent on three different views, we perform 3-dimensional rotation [21] along *X*, *Y*, and *Z*-axes for making our system view independent. The common form of 3-dimensional view rotation are defined in equations (4)-(6):

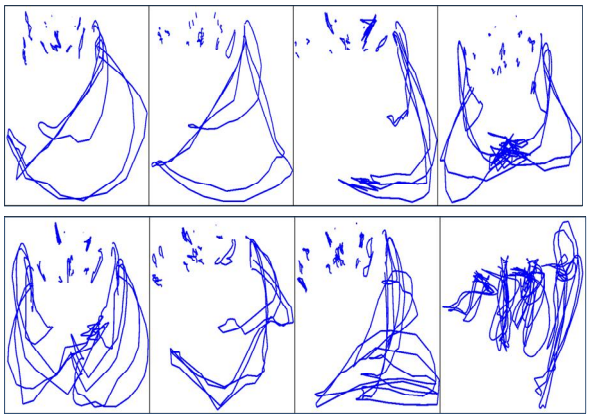
$$R_x(\theta) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & -\sin\theta \\ 0 & \sin\theta & \cos\theta \end{bmatrix} \tag{4}$$



(a) Mapping of eight action along *XY*-axes (front view)



(b) Mapping of eight action along *YZ*-axes (side view)



(c) Mapping of eight action along *ZX*-axes (top view)

그림 5. 엔드-투-엔드 스켈레톤 조인트 매핑
Fig. 5. End-to-end skeleton joints mapping.

$$R_y(\theta)=\begin{bmatrix}\cos\theta & 0 & \sin\theta\\ 0 & 1 & 0\\ -\sin\theta & 0 & \cos\theta\end{bmatrix} \quad (5)$$

and

$$R_z(\theta)=\begin{bmatrix}\cos\theta & -\sin\theta & 0\\ \sin\theta & \cos\theta & 0\\ 0 & 0 & 1\end{bmatrix} \quad (6)$$

where $R_{axes}(\theta)$ represent the rotation of data along *X*, *Y*, and *Z*-axes with angle θ.

3-3 Deep Convolution Neural Network

For the discrimination among the action classes, we use a modified version of pre-trained AlexNet. We integrate one additional block containing convolution, ReLU, pooling and replace the input and out layers for fitting our classification purpose. There are eight blocks in which first two blocks having convolution, ReLU, normalization following by pooling layers. The third, fourth and fifth blocks consisting of convolution and ReLU layers. The sixth block integrates convolution, ReLU and pooling layers. The seventh block consists of fully connected, ReLU, normalization following by dropout and the last block combines a fully connected and softmax layers. The details of the network are shown in Fig. 6.

The size of the input to the network is 227×227×3 for all the views along *XY*, *YZ*, and *ZX*-axes. The input is first passed through a convolution layer having 11×11×3 kernel, 96 output filters, same padding, and 4 stride that produces output of size 55 ×55×96 followed by ReLU and batch normalization. The output is passed through a max-pooling layer having 3×3×3 kernel, and 2 stride that produces output of size 27×27×96. The learnable parameters weight, and bias for first convolution layer are 11×11×3×96, and 1×1×96 respectively. The above operations continue two times and generate the output of size 13 ×13×256. Then the generated output passes through the three blocks intended to do the same operations of convolution and ReLU which outcomes the same size as the input of 13×13× 256. The sixth block takes the results from the fifth block and performs convolution, ReLU and max-pooling operations. It provides the output of size 6×6×128. Then the generated output of size 6×6×128 passes through a fully connected layer followed by a ReLU, batch normalization, and dropout layers of 50% that produces output of size 1×1×128. The second fully connected layer produces the final the outputs of size 1×1×27 followed by a softmax and a classification layer [22].

Ⅳ. Experimental Results and Performance Analysis

This section conducts the details of dataset, experimental

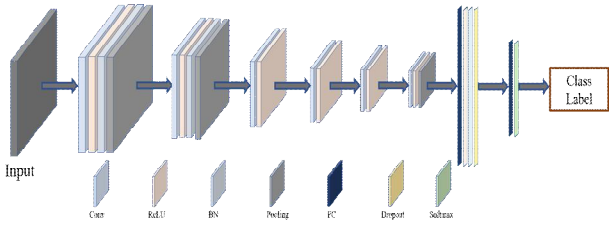


그림 6. DCNN 네트워크의 구조
Fig. 6. The DCNN Network Architecture.

setting, performance evaluation, and comparison.

4-1 Dataset

We use the dataset UTD-MHAD of the 3D skeleton dataset provided by ESSP Laboratory at the University of Texas at Dallas [20]. The dataset is captured using a Microsoft Kinect sensor in an indoor environment. It contains 27 different classes of skeleton data in which each frame having 20 joints along *X*, *Y*, and *Z*-axes. The 27 classes of actions are performed by 8 different subjects including 4-females and 4-males. Each class has 32 videos except three of them that are corrupted making a total of 861 videos of skeleton data. Most of the action is performed by hands like swipe left, swipe right, wave, clap, throw, arm cross, basketball shoot, draw x, draw circle, draw triangle, bowling, boxing, baseball swing, tennis swing, arm curl, tennis serve, push, knock, catch, pickup, and throw. Some of the actions are also captured by leg such as Jogging, walking, lunging. There are only two actions which are acted by the full body. Each person repeated each action 4 times in every 27 classes.

We evaluate the method through three different experiments along *XY*, *YZ*, and *ZX*-axes and then fuse all of them by concatenating 3 views in order to get better results. The experimental results are shown in terms of accuracy given by the following equation (7):

$$Accuracy(\%) = \frac{M}{N} \times 100 \tag{7}$$

where M is the total correctly predicted observations and N is the total number of observations. The details of implementation are listed in Table 1.

We initially assign the learning rate of 0.001 which decreases into the half after every fifth epoch. We train our model until the completion of the twenty epoch for getting the desired results.

4-2 Performance Evaluation and Comparison

Most of the actions are more classifiable when mapping along *XY* and *ZX*-axes rather than *YZ*-axes. Thus we get higher accuracy while training and testing the data for *XY* and *ZX*-axes. The results by combining all three different views are much greater than individual views. We get the training results above 95% for all the cases. The classification accuracies for testing the trained model are given in Table 2.

By observing the results in Table 2, it is clear that we get 95.76% classification accuracy for mapping joints along *XY*-axes which is larger in compared to the results of *YZ* and *ZX*-axes (92.48% and 94.10% respectively). The concatenation of *XY*, *YZ* and *ZX*-axes have strong tendency to classify the actions that shows about 97.45% accuracy.

For establishing the effectiveness and robustness of the proposed method, we compare the results with four related methods described in section 2. The results are shown in Table 3.

As described in [19], the classification accuracy was 93.48% that is the highest result using the UTD-MHAD skeleton dataset. The accuracies of the rest of the methods are less than 90%. We get the highest accuracy among all of the existing systems as demonstrated earlier section. Thus, it can be concluded that our proposed method outperforms among the existing systems.

Ⅴ. Conclusion

표 1. 파라미터 세팅
Table 1. Parameters setting.

Parameters	Values
Number of epoch	20
Initial learning rate	0.001
LearningRateDropFactor	0.5
LearningRateDropPeriod	5

표 2. 분류 결과
Table 2. Classification results.

Mapping	Accuracy
Along *XY*-axes	95.76%
Along *YZ*-axes	92.48%
Along *ZX*-axes	94.10%
Fusion	97.45%

표 3. 성능 비교
Table 3. Performance comparison.

Methods	Accuracy
JTM [16]	85.81%
SOS [17]	86.97%
JDMs [18]	88.10%
BiGRU [19]	93.48%
Ours	97.45%

We introduce a noble method for action recognition based on skeleton data using a deep convolution neural network. First, the mapping of the skeleton joints is done along the temporal direction and then discriminates using the DCNN for deciding the final class. We perform experiments on three different views along *XY*, *YZ*, and *ZX*-axes and then fused all of them to generate the final results. The proposed method outperforms over the existing systems in case of side, front, top and fused results.

Acknowledgments

This research was supported by the GRRC program of Gyeonggi province [GRRC Aviation 2017-B04, Development of Interactive VR Player and Service with Space Media Convergence].

References

[1] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston: MC, pp. 1110–1118, 2015.

[2] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *IEEE Conference on Computer Vision and Pattern Recognition*, Columbus: OH, pp. 804–811, 2014.

[3] V. S. Kulkarni, and S. D. Lokhande, "Appearance based recognition of american sign language using gesture segmentation," *International Journal on Computer Science and Engineering*, No. 3, pp. 560-565, 2010.

[4] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Breckenridge: CO, pp. 65–72, 2005.

[5] D. Wu, and L. Shao, "Silhouette analysis-based action recognition via exploiting human poses," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 23, No. 2, pp. 236-243, 2012.

[6] M. Ahmad, and S. W Lee, "HMM-based human action recognition using multiview image sequences," in *18th International Conference on Pattern Recognition (ICPR'06)*, Hong Kong, pp. 263-266, 2006.

[7] L. Xia, C. C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Providence: IR, pp. 20-27, 2012.

[8] J. Luo, W. Wang, and H. Qi, "Spatio-temporal feature extraction and representation for RGB-D human action recognition," *Pattern Recognition Letters*, Vol. 50, pp. 139-148, 2014.

[9] V. Megavannan, B. Agarwal, and R. V. Babu, "Human action recognition using depth maps," in *2012 International Conference on Signal Processing and Communications* (*SPCOM*), Piscataway: NJ, pp. 1-5, 2012.

[10] J. Trelinski, and B. Kwolek, "Convolutional neural network-based action recognition on depth maps," in *International Conference on Computer Vision and Graphics,* Warsaw: Poland, pp. 209-221, 2018.

[11] P. Wang, W. Li, Z. Gao, J. Zhang, C. Tang, and P.O. Ogunbona, "Action recognition from depth maps using deep convolutional neural networks," *IEEE Transactions on Human-Machine Systems*, Vol. 46, No. 4, pp. 498-509, 2015.

[12] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems*, Montreal: Canada, pp. 568–576, 2014.

[13] C. Li, Q. Zhong, D. Xie, and S. Pu, "Skeleton-based action recognition with convolutional neural networks," in *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW),* Hong Kong, pp. 597-600, 2017.

[14] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban, "Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations," in *Twenty-Third International Joint Conference on Artificial Intelligence*, Beijing: China, pp. 2466-2472, 2013.

[15] Y. Du, Y. Fu, and L. Wang, "Skeleton based action recognition with convolutional neural network," *IEEE 2015 3rd IAPR Asian Conference on Pattern Recognition* (*ACPR*), Kuala Lumpur: Malaysia, pp. 579-583, 2015.

[16] P. Wang, Z. Li, Y. Hou, and W. Li, "Action recognition based on joint trajectory maps using convolutional neural networks," *in Proceedings of the 24th ACM International Conference on ACM Multimedia,* Amsterdam: Netherlands, pp. 102-106, 2016.

[17] Y. Hou, Z. Li, P. Wang, and W. Li, "Skeleton optical spectra-based action recognition using convolutional neural networks," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 28, No. 3, pp. 807-811, 2016.

[18] C. Li, Y. Hou, P. Wang, and W. Li, "Joint distance

maps-based action recognition with convolutional neural networks," *IEEE Signal Processing Letters*, Vol. 24, No. 5, pp. 624-628, 2017.

[19] J. Imran, and B. Raman, "Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1-20, 2019.

[20] UTD-MHAD skeleton dataset, University of Texas at Dalas, [Internet]. Available: https://personal.utdallas.edu/~kehtar/UTD-MHAD.html

[21] C. Shorten, and TM. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, Vol. 6, No. 1, pp. 60, 2019.

[22] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, Lake Tahoe: NV, pp. 1097-1105, 2012.

타 스 님 (Nusrat Tasnim)

2017년 12월 : University of Chittagong(Bangladesh), Computer Science & Engineering (공학사)
2020년 2월 : 한국항공대학교 항공전자정보공학과 (공학석사)
2020년 ~ 현재 : 한국항공대학교 항공전자정보공학과 박사과정
※관심분야 : 컴퓨터비전, 멀티미디어

백 중 환 (Joong-Hwan Baek)

1981년 2월 : 한국항공대학교 항공통신공학 (공학사)
1987년 7월 : 오클라호마주립대학원 전기 및 컴퓨터공학 (공학석사)
1991년 7월 : 오클라호마주립대학원 전기 및 컴퓨터공학 (공학박사)
1992년 ~ 현재 : 한국항공대학교 항공전자정보공학부 교수
※관심분야 : 영상처리, 패턴인식, 멀티미디어, 가상현실