한국정보전자통신기술학회논문지(jkiiect)'21-8, Vol.14 No.4
논문 21-14-04-314 http://dx.doi.org/10.17661/jkiiect.2021.14.4.314

# A Proposal of Shuffle Graph Convolutional Network for Skeleton-based Action Recognition

Sungjun Jang, Han Byeol Bae, HeanSung Lee and Sangyoun Lee*

**Abstract** Skeleton-based action recognition has attracted considerable attention in human action recognition. Recent methods for skeleton-based action recognition employ spatiotemporal graph convolutional networks (GCNs) and have remarkable performance. However, most of them have heavy computational complexity for robust action recognition. To solve this problem, we propose a shuffle graph convolutional network (SGCN) which is a lightweight graph convolutional network using pointwise group convolution rather than pointwise convolution to reduce computational cost. Our SGCN is composed of spatial and temporal GCN. The spatial shuffle GCN contains pointwise group convolution and part shuffle module which enhances local and global information between correlated joints. In addition, the temporal shuffle GCN contains depthwise convolution to maintain a large receptive field. Our model achieves comparable performance with lowest computational cost and exceeds the performance of baseline at 0.3% and 1.2% on NTU RGB+D and NTU RGB+D 120 datasets, respectively.

**Key Words** : Convolutional neural network, Graph convolutional network, Lightweight neural network, Pointwise group convolution, Skeleton-based action recognition

## 1. Introduction

Human action recognition is used in various real-world applications, such as human robot interaction and video surveillance[1,2]. Especially, skeleton-based action recognition has attracted extensive attention since it can extract the robust features regardless of background interference, scale changes, or illumination conditions compared with RGB and video data [3, 4, 5, 6, 7, 8].

Conventional methods for skeleton-based action recognition considered each human joint as independent element, and designed handcrafted features to represent skeleton. Following the development of deep learning, researchers construct the skeleton data as a sequence of coordinate vectors or a fixed 2D grid that is fed into recurrent neural network (RNN)[10, 11, 12, 13] or convolution neural network (CNN) [14, 15, 16, 17, 18]. However, these approaches cannot represent corresponded joint dependencies on spatiotemporal dimension.

Yan et al. [3] first apply graph convolutional networks(GCNs), constructing a spatiotemporal graph that considers 3D coordinate body keypoints as nodes and their natural connectivities as edges. Their spatiotemporal graph convolutional network (ST-GCN) achieves better performance compared with previous methods. With various models based on ST-GCN [7, 19, 6, 4, 5], recent methods have achieved high performance by designing multi-scale structures or introducing various

*Corresponding Author : Department of Electrical and Electronic Engineering Yonsei University(syleee@yonsei.ac.kr)
Received August 13, 2021 Revised August 14, 2021 Accepted August 20, 2021

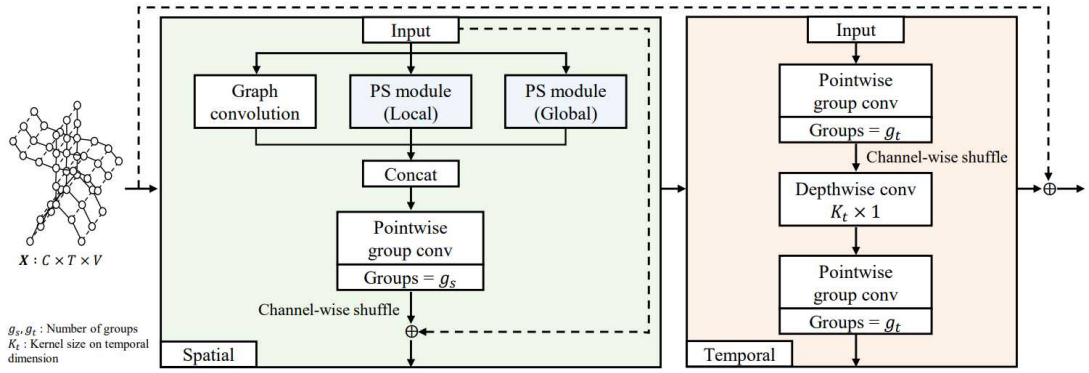


Fig. 1. Overall architecture of our shuffle graph convolutional network (SGCN). The spatial graph convolutional network contains graph convolution and part shuffle (PS) modules, and the temporal graph convolutional network has large receptive field to obtain temporal information with low computational cost.

modules to enhance network ability. However, most of them have heavy computational cost because they construct large graphs or build multiple network. To solve this shortcoming, shift-GCN [8] proposes shift graph convolutional network using shift operation and pointwise convolution.

In this work, we propose a lightweight shuffle graph convolutional network (SGCN). Our network consists of spatial and temporal shuffle GCNs, which is inspired by shufflenet [20]. The proposed spatial shuffle GCN contains graph convolution using a single adjacency matrix and part shuffle (PS) module that can effectively capture local and global joint connectivities. The proposed temporal shuffle GCN achieves better performance with lower computational cost while maintaining the same receptive field size as standard convolution approaches.

The main contributions are as follows: We propose a lightweight spatiotemporal GCN using pointwise group convolution and depthwise convolution. Our model has flexibility to reduce computational cost with the various number of groups.

The proposed PS module enhances feature representation using local and global joint connectivities and achieves better performance to predict action labels.

## 2. PROPOSED METHOD

### 2.1. Overview

The goal of the skeleton based action recognition is to predict the action label for given data. The data is a sequence of frames, where each frame contains joint coordinates. The spatiotemporal GCN extracts the features from input data. We adopt the same two-stream framework as [4], which fuses scores by training joint and bone data in the identical network. Bone data is represented by a vector pointing to the target joint from the source joint, and contains length and direction information. Softmax scores from two stream models are then summed to obtain final scores and predict the action label.

Skeleton data in this work is denoted as $G=(V,E)$ where $V$ is the joints as a set of nodes and $E$ is the bones as a set of edges. The

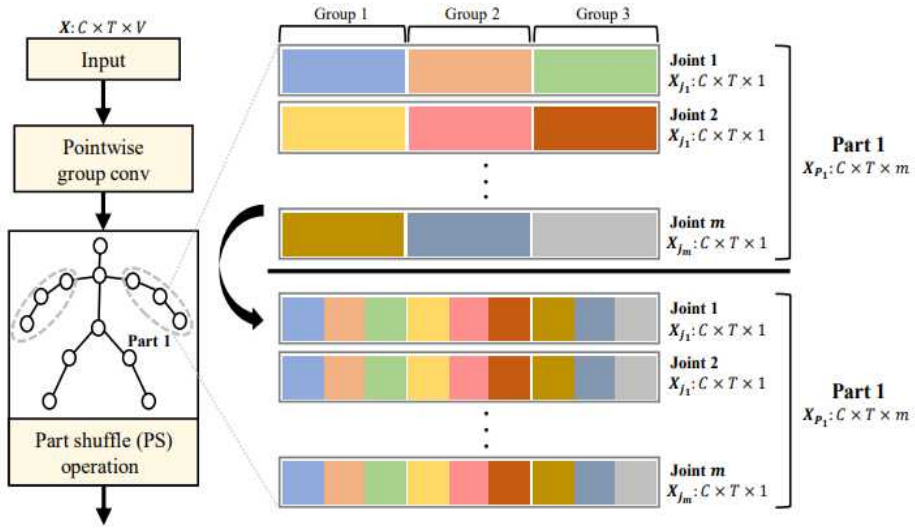


Fig. 2. Illustration of the part shuffle (PS) module. By implementing part shuffle operation locally and globally, the features of each joint have a large receptive field.

data represents the skeleton sequence, so we denote input data as $X \in R^{C \times T \times V}$, where $C$, $T$, and $V$ are the number of channels, frames, and joints, respectively.

We use a group convolution, which divides the channel axis from standard convolution to group-wise, and we represent the groups of spatial and temporal networks as variables $g_s$ and $g_t$, respectively.

## 2.2. Spatial Shuffle Graph Convolutional Network

Fig. 1. shows the proposed shuffle graph convolutional network containing three branches on the spatial GCN. First branch is graph convolutional network represented as:

$$X_{t_{out}} = A' X_t W, \qquad (1)$$

where $W$ is the pointwise convolution weights. $X_t$ is the features extracted from frame $t$.

$A' = A + I$ where $A$ is a learnable adjacency matrix, and $I$ is the identity matrix. The remaining branches are local and global PS modules that combine the joints into a semantic part to obtain local or global information. The outputs from each branch are integrated by the group convolution, where $g_s$ is the number of groups. We then shuffle the features on channel dimension to enhance representation capability. Our spatial shuffle GCN can take various semantic connectivities between joints with low computational cost.

## 2.3. Part Shuffle (PS) Module

In this work, a part means a set of joints that are divided based on semantic connectivities. The PS module contains local and global modules and extracts the features from each semantic part. For the local PS module, we define the semantic parts as 'left arm', 'right arm', 'left leg', 'right leg', and 'head'.

As shown in Fig. 2., each part has the features of $R^{C \times T \times m}$, where $m$ is the number

of joints. Each joint of $X_{j_i} \in R^{C \times T \times 1} (i = 1, 2, \ldots, V)$ is divided into channel groups by the group convolution.

For simplicity, the number of group is set to 3. We propose a part shuffle(PS) operation to take a large receptive field by simultaneously shuffling features on the channel and joint dimensions. After PS operation, each joint takes the features of other joints that belong to the same part, to capture the local information on spatial dimension. For global PS module, we assume all the joints are the same part and shuffle features from all joints by PS operation.

### 2.4. Temporal Shuffle Graph Convolutional Network

Since action data is a sequence of frames, modeling temporal network is necessary to extract the temporal information. Many existing methods use the $K_t \times 1$ convolution to design temporal GCN, where $K_t$ is the kernel size on temporal dimension. Our temporal shuffle GCN is composed of pointwise group convolution and depthwise convolution layers to reduce the computational cost. To enhance the feature representation, we apply channel-wise shuffle operation after the first pointwise group convolution.

For comparison of FLOPs, we denote the input features as $F \in R^{C \times H \times W}$, where $C$ is the number of channels and $H \times W$ is the feature map size. When the kernel $K \in R^{K_t \times 1 \times C \times C'}$ is used in standard convolution, its FLOPs is calculated as:

$$HW \times K_t \times C \times C'.$$

Whereas, the FLOPs of our temporal model is calculated as:

$$HW \times (2C/g_t + K_t C'),$$

where $g_t$ is the number of groups in the group convolution.

### 2.5. Implementation Details

Tested on a single Titan RTX GPU, our model achieves a classification accuracy of 88.8% on the NTU-RGB+D. The overall framework of our SGCN includes 8 layer except for a fully-connected layer, where each layer contains spatial and temporal shuffle GCNs. We use cross-entropy (CE) loss function and SGD optimizer with momentum of 0.9. our models are trained for 120 epochs with the initial learning rate of 0.1. The learning rate is divided by 10 at epoch 60, 80, and 100. The batch size is 32 and the weight decay is set to 0.0001.

For NTU RGB+D and 120 [10, 25], the maximum number of frames is 300 in each sample. For samples less than 300 frames, we repeat the sample until it reaches 300 frames. Each sample contains at most two people. If there are less than two people in the sample, we pad the body of the second person with 0. Before training, inputs are pre-processed with normalization and translation, similar to existing works [4].

## 3. EXPERIMENT

### 3.1 Datasets

NTU RGB+D [10] is the most commonly used dataset in action recognition, containing

Table 1. Comparisons of the Top-1 accuracy (%) and FLOPs(G) with state-of-the-art methods on the NTU-RGB+D.

| Methods | xsub | xview( | FLOPs( |
|---|---|---|---|

| | (%) | %) | G) |
|---|---|---|---|
| Lie Group [9] | 50.1 | 52.8 | - |
| Deep LSTM [10] | 60.7 | 67.3 | - |
| ST-LSTM [11] | 69.2 | 77.7 | - |
| STA-LSTM [12] | 73.4 | 81.2 | - |
| Ind-RNN [13] | 81.8 | 88.0 | - |
| 2s-3DCNN [14] | 66.8 | 72.6 | - |
| TCN [15] | 74.3 | 83.1 | - |
| 3scale ResNet152 [16] | 85.0 | 92.3 | - |
| HCN [17] | 86.5 | 91.1 | - |
| TS-SAN [18] | 87.2 | 92.7 | - |
| ST-GCN [3] | 81.5 | 88.3 | - |
| 2s-ASGCN [7] | 86.8 | 94.2 | 27.0 |
| 2s-AGCN [4] | 88.5 | 95.1 | 35.8 |
| 2s-AGCLSTM [5] | 89.2 | 95.0 | 54.4 |
| 2s-DGNN [6] | 89.2 | 95.5 | 63.4 |
| 2s-ShiftGCN [8] | **89.7** | **96.0** | 5.0 |
| 2s-SGCN (ours) | 88.8 | 94.7 | **3.3** |

56,880 video clips and 60 classes. Video clips have 25 joint data of the body obtained from a kinetic sensor. It provides information about a total of 40 people and data taken from three different camera views. The author recommends two benchmarks to evaluate recognition performance. 1) Cross-subject (xsub) is divided into 40,320 training sets and 16,560 validation sets, consisting of different subjects. 2) Cross-view (xview) consists of 37,920 training sets, including two camera views at different angles. The rest of clips are used as validation set.

NTU RGB+D 120 [25] is an extended version of NTU RGB+D. It provides an additional 57,000 video clips and 60 added classes, consisting of 110,000 clips for a total of 120 classes from 106 different people and 32 different camera setup. Cross subject uses 54,000 clips for 53 different subjects as training, while the remaining clips are used as validation sets. The author recommends to replace the cross-view of NTU RGB+D with

Table 2. Comparisons of the Top-1 accuracy (%) and FLOPs(G) with state-of-the-art methods on the NTU-RGB+D 120.

| Methods | xsub (%) | xset (%) | FLOPs (G) |
|---|---|---|---|
| ST-LSTM [11] | 55.7 | 57.9 | - |
| GCA-LSTM [21] | 61.2 | 63.3 | - |
| RotClips+CNN [22] | 62.2 | 61.8 | - |
| Body Pose Evolution Map [23] | 66.9 | 64.6 | - |
| Skelemotion [24] | 67.7 | 66.9 | - |
| 2s-AGCN [4] | 82.9 | 84.9 | 35.8 |
| 2s-ShiftGCN [8] | **85.3** | **86.6** | 5.0 |
| 2s-SGCN (ours) | 84.1 | 85.5 | **3.3** |

cross-setup. Cross setup (xset) uses 54,000 clips containing half setup for training, while the rest of clips are used as validation sets.

### 3.2 Evaluation

To validate our proposed method, we compare our SGCN with state-of-the-art methods on NTU RGB+D and NTU RGB+D 120 datasets. We include handcrafted feature-based methods, RNN-based methods, CNN-based methods, and GCN-based methods. We divide them into horizontal lines. Since many state-of-the-art GCN-based methods utilize multi-stream fusion, and we also adopt the same approach to ensure fair comparison.

On NTU RGB+D dataset, our model is evaluated on cross-subject(xsub) and cross-view(xview) benchmarks in Table 1. The GCN-based methods generally perform better than RNN and CNN based-methods. Our model exceeds higher accuracy than 2s-AGCN [4] with

Table 3. Comparisons between standard GCNs and our spatiotemporal shuffle GCN on the NTU RGB+D.

| Spatial | Temporal | xsub(%) | FLOPs (G) |
|---|---|---|---|

| | | | |
|---|---|---|---|
| Standard | Standard | 88.5 | 35.8 |
| Standard | Shuffle | **89.4** | 14.9 |
| Shuffle | Standard | 88.8 | 24.2 |
| Shuffle (w/o PS) | Shuffle | 88.5 | **3.3** |
| Shuffle | Shuffle | 88.8 | **3.3** |

10.8× less computational cost, and achieves comparable performance with lowest computational cost compared to other state-of-the-art methods.

Table 2 shows the performance of cross-subject(xsub) and cross-setup(xset) benchmarks for the various models on the NTU RGB+D 120 dataset, with similar comparable performance and efficiency of our model. Although there is some trade-off between accuracy and computational cost, our model is able to be less affected and get more flexibility to reduce computational cost by adjusting the number of $g_s$ and $g_t$.

### 3.3 Ablation Studies

We conduct ablation studies to demonstrate our work. In Table 3, we show the effectiveness and efficiency of our spatial and temporal networks by comparing them with baseline model (2s-AGCN [4]). We describe the proposed method as shuffle and baseline method as standard. When each spatial and temporal shuffle GCN is applied, we achieve better performance with much less computational cost than baseline. Our final model, including both spatial and temporal shuffle GCNs, exceeds the performance of baseline at 0.3% with lower computational cost. Unlike other models, it allows flexibly to reduce computational cost by using the shuffle method and mitigates trade-off of the performance through PS opeartion that gives correlation between shuffled features and the parts of skeleton.

Table 4. Recognition accuracy and computational cost with various number of groups in the spatiotemporal shuffle graph convolutional network on the NTU RGB+D.

| Methods | $g_s$ | $g_t$ | xsub(%) | FLOPs(G) |
|---|---|---|---|---|
| SGCN | 3 | 3 | **88.8** | 3.3 |
| | 3 | 5 | 88.2 | 3.2 |
| | 5 | 3 | 88.1 | 3.0 |
| | 5 | 5 | 88.3 | **2.9** |

Table 3 also shows that the part shuffle operation is beneficial for spatial shuffle GCN and deleting this operation will harm the performance.

In Table 4, we show our SGCN can flexibly reduce the computational cost with various number of groups. We evaluate by changing $g_s$ and $g_t$ to 3 and 5 on the spatiotemporal shuffle GCN. It shows the best performance of 88.8% when both $g_s$ and $g_t$ are set to 3. When $g_s$ and $g_t$ are different, the features of spatial and temporal shuffle GCN have an unbalanced receptive field. Thus, this causes degradation of representation capability and lower performance than the same number of groups.

In Table 5, we show the effectiveness of PS module. The result shows that given each branch, both local and global PS modules are important, which also proves importance of connectivities between joints.

Table 5. Comparisons of the accuracy obtained by the local and global PS module on the NTU RGB+D.

| Methods | | | xsub(%) |
|---|---|---|---|
| GCN | Local | Global | |
| ✓ | | | 88.4 |
| ✓ | ✓ | | 88.5 |
| ✓ | | ✓ | 88.4 |
| ✓ | ✓ | ✓ | **88.8** |

## 4. CONCLUSION

In this work, we have proposed a shuffle graph convolutional network (SGCN) for skeleton-based action recognition which contains spatial and temporal shuffle GCN. Our SGCN has more flexibility to reduce computational cost which is composed of pointwise group convolution rather than pointwise convolution.

The spatial shuffle GCN can take the various joint connectivities by implementing PS module. The temporal shuffle GCN achieves better performance than standard convolution with much less computational cost and can take a large receptive field.

The final model achieves comparable performance to current state-of-the-art method with lowest computational cost on NTU RGB+D and NTU RGB+D 120 datasets.

## REFERENCES

[1] Utkarsh Gaur, Yingying Zhu, Bi Song, and A Roy-Chowdhury,"A "string of feature graphs" model for recognition of complex activities in natural videos," in 2011 International Conference on Computer Vision. IEEE, 2011, pp. 2595-2602.

[2] Zoran Duric, Wayne D Gray, Ric Heishman, Fayin Li, Azriel Rosenfeld, Michael J Schoelles, Christian Schunn, and Harry Wechsler, "Integrating perceptual and cognitive modeling for adaptive and intelligent human-computer interaction," Proceedings of the IEEE, vol. 90, no. 7, pp. 1272-1289, 2002.

[3] Sijie Yan, Yuanjun Xiong, and Dahua Lin, "Spatial temporal graph convolutional networks for skeleton-based action recognition," in Proceedings of the AAAI Conference on Artificial Intelligence, 2018, vol. 32.

[4] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, "Twostream adaptive graph convolutional networks for skeletonbased action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 12026-12035.

[5] Chenyang Si, Wentao Chen, Wei Wang, Liang Wang, and Tieniu Tan, "An attention enhanced graph convolutional lstm network for skeleton-based action recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 1227-1236.

[6] Lei Shi, Yifan Zhang, Jian Cheng, and Hanqing Lu, "Skeletonbased action recognition with directed graph neural networks," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7912-7921.

[7] Maosen Li, Siheng Chen, Xu Chen, Ya Zhang, Yanfeng Wang, and Qi Tian, "Actional-structural graph convolutional networks for skeleton-based action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3595-3603.

[8] Ke Cheng, Yifan Zhang, Xiangyu He, Weihan Chen, Jian Cheng, and Hanqing Lu, "Skeleton-based action recognition with shift graph convolutional network," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 183-192.

[9] Raviteja Vemulapalli, Felipe Arrate, and Rama Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 588-595.

[10] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang, "Ntu rgb+ d: A large scale dataset for 3d human activity analysis," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1010-1019.

[11] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang, "Spatiotemporal lstm with trust gates for 3d human action recognition," in

European conference on computer vision. Springer, 2016, pp. 816-833.

[12] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," in Proceedings of the AAAI conference on artificial intelligence, 2017, vol. 31.

[13] Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao, "Independently recurrent neural network (indrnn): Building a longer and deeper rnn," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 5457-5466.

[14] Hong Liu, Juanhui Tu, and Mengyuan Liu, "Two-stream 3d convolutional neural network for skeleton-based action recognition," arXiv preprint arXiv: 1705. 08106, 2017.

[15] Tae Soo Kim and Austin Reiter, "Interpretable 3d human action analysis with temporal convolutional networks," in 2017 IEEE conference on computer vision and pattern recognition workshops (CVPRW). IEEE, 2017, pp. 1623-1631.

[16] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu, "Skeletonbased action recognition with convolutional neural networks," in 2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2017, pp. 597-600.

[17] Chao Li, Qiaoyong Zhong, Di Xie, and Shiliang Pu, "Cooccurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation," arXiv preprint arXiv: 1804. 06055, 2018.

[18] Sangwoo Cho, Muhammad Maqbool, Fei Liu, and Hassan Foroosh, "Self-attention network for skeleton-based human action recognition," in The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 635-644.

[19] Ziyu Liu, Hongwen Zhang, Zhenghao Chen, Zhiyong Wang, and Wanli Ouyang, "Disentangling and unifying graph convolutions for skeleton-based action recognition," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 143-152.

[20] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun, "Shufflenet: An extremely efficient convolutional neural network for mobile devices," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6848-6856.

[21] Jun Liu, Gang Wang, Ling-Yu Duan, Kamila Abdiyeva, and Alex C Kot, "Skeleton-based human action recognition with global context-aware attention lstm networks," IEEE Transactions on Image Processing, vol. 27, no. 4, pp. 1586-1599, 2017.

[22] Qiuhong Ke, Mohammed Bennamoun, Senjian An, Ferdous Sohel, and Farid Boussaid, "Learning clip representations for skeleton-based 3d action recognition," IEEE Transactions on Image Processing, vol. 27, no. 6, pp. 2842-2855, 2018.

[23] Mengyuan Liu and Junsong Yuan, "Recognizing human actions as the evolution of pose estimation maps," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1159-1168.

[24] Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot, "Global context-aware attention lstm networks for 3d action recognition," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1647-1656.

[25] Jun Liu, Amir Shahroudy, Mauricio Lisboa Perez, Gang Wang, Ling-Yu Duan, and Alex Kot Chichung, "Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding," IEEE transactions on pattern analysis and machine intelligence, 2019.

## 저자약력


**장 성 준(Sungjun Jang) [정회원]**

- 2019년 광운대학교 전자공학과 학사 졸업
- 현재 : 연세대학교 석박 통합과정

〈관심분야〉 행동 인식, 얼굴 인식, 이미지 분할, 경량화 네트워크 등

**배 한 별(Han Byeol Bae) [정회원]**

- 2010년 연세대학교 컴퓨터정보통신공학과, 전기전자공학과 학사 졸업
- 2015년 연세대학교 생체인식 협동과정 석사 졸업
- 2020년 연세대학교 전기전자공학과 박사 졸업
- 현재 : 연세대학교 박사 후 연구원

〈관심분야〉 얼굴 인식, 패턴인식, 이미지 변환, 이미지 분류 등

**이 현 성(HeanSung Lee) [정회원]**

- 2018년 한국항공대학교 항공전자정보공학과 학사 졸업
- 현재 : 연세대학교 석박 통합 과정

〈관심분야〉 다중 물체 추적, 물체 재인식, 물체 검출, 이미지 분할 등

**이 상 윤(Sangyoun Lee) [정회원]**

- 1987년 연세대학교 전기전자공학과 학사 졸업
- 1989년 연세대학교 전기전자공학과 석사 졸업
- 1999년 조지아공과대학 전기 및 컴퓨터 공학과 박사
- 현재 : 연세대학교 전기전자공학 교수

〈관심분야〉 얼굴인식, 패턴인식, 비디오 코덱 등