Skeleton-Based Human Action Recognition

송 재 훈

Vision & Display Systems Lab.

Dept. of Electronic Engineering, Sogang University

Outline

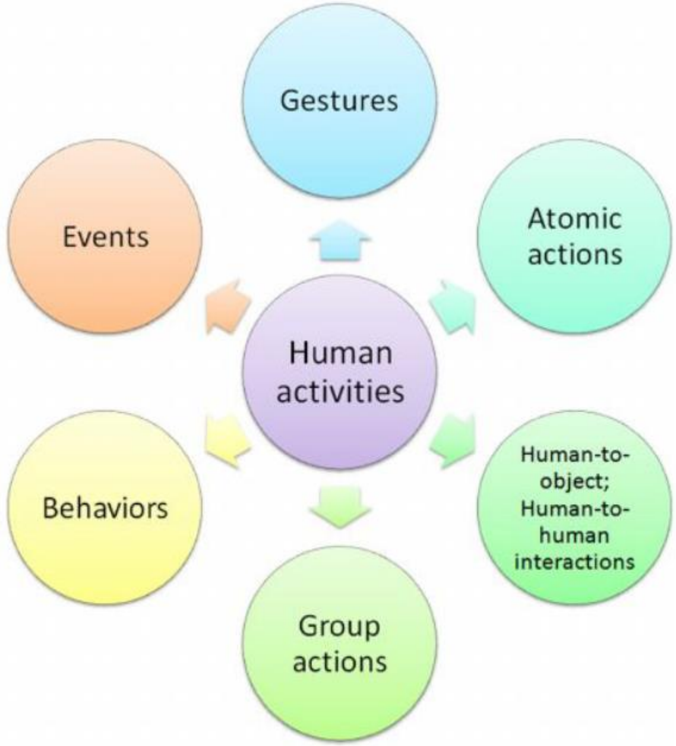
- Human Action Recognition 이란
- Introduction
- Application
- Action Recognition from Skeletons
- Flow of Action Recognition
- Datasets
- ST-GCN paper
- ST-GCN 이후 research

Introduction

- Human Action Recognition
    - 사람의 행동을 분류하는 작업
        - 사람의 모든 행동은 목적을 달성하기 위해 수행된다
        - Machines는 이를 배우고 이해할 수 있어야 함

Introduction

- Levels of visual source understanding
  - Object-level understanding
  - Tracking-level understanding
  - Pose-level understanding
  - Activity-level understanding

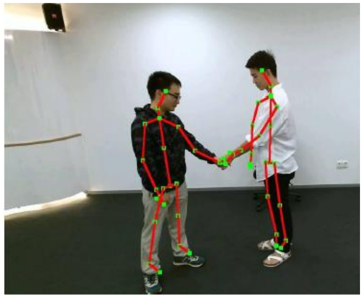


Application

- Many potential applications of action recognition systems
  - Video Retrieval
  - Video Surveillance
  - Health Care
  - Human-Computer Interaction
  - Entertainment Industry
  - …..

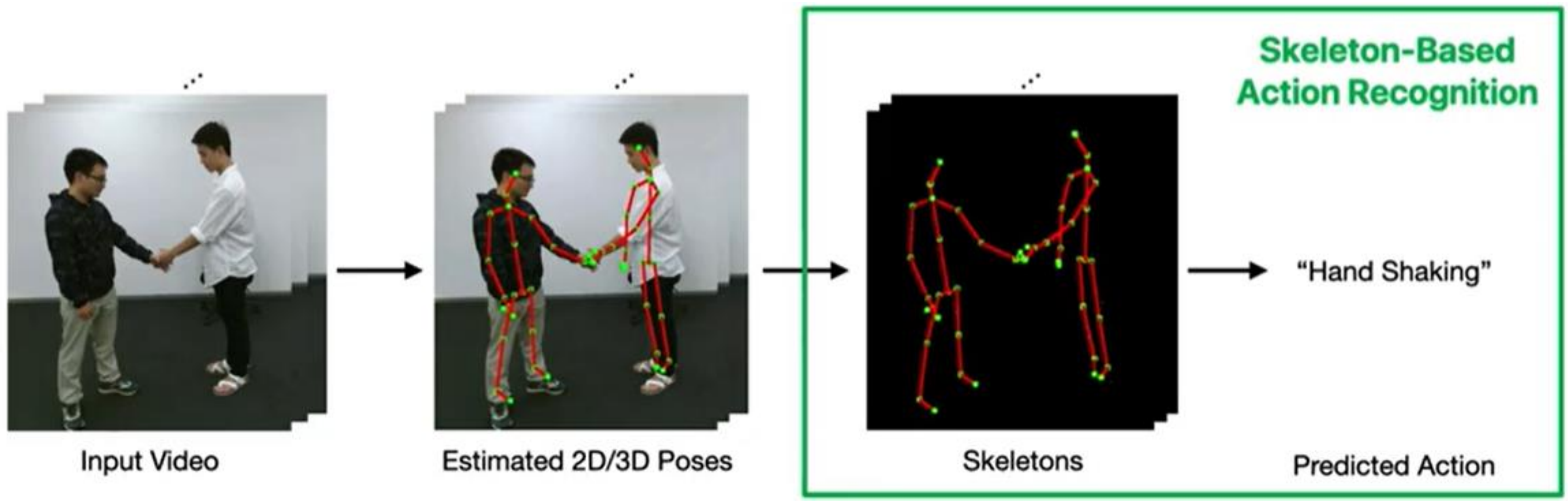



- HW/SW의 발전으로 실생활에 적용 가능

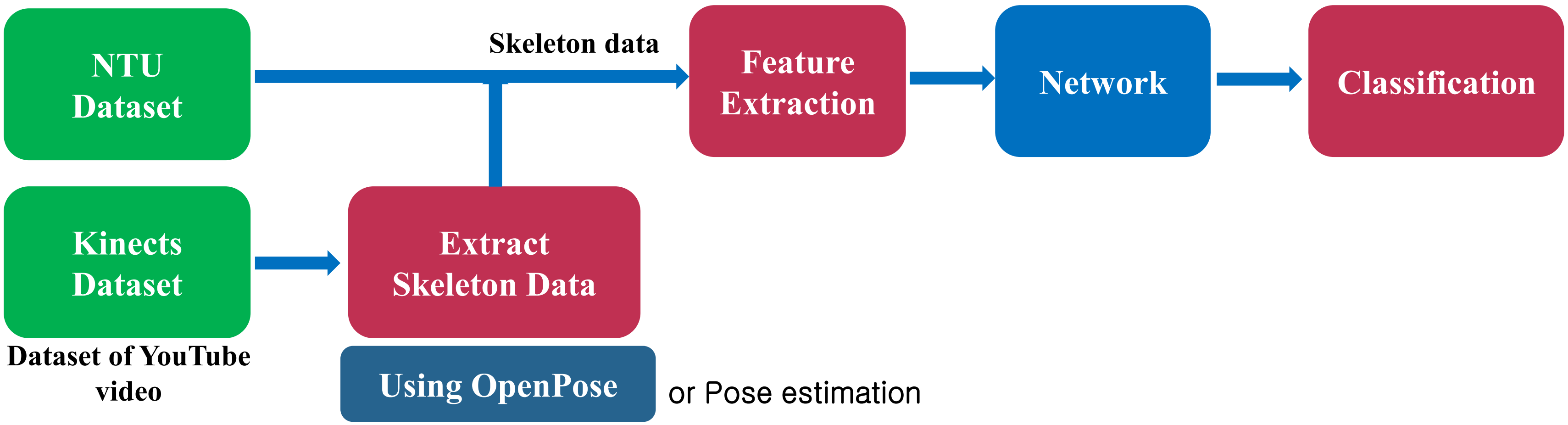

Action Recognition from Skeletons

- Human actions can be efficiently represented by skeletons
- Free of background clutter / lighting conditions / clothing variations

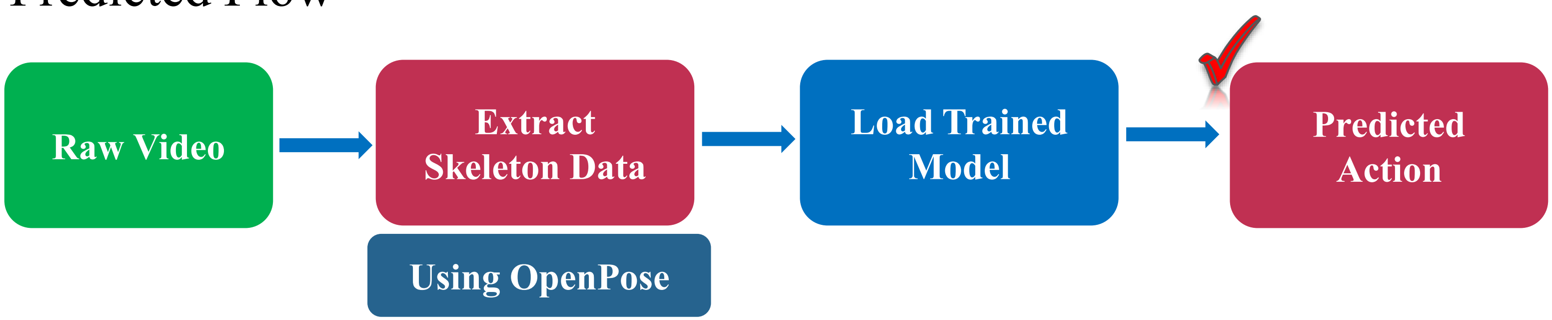


Flow of Action Recognition

- Training Flow

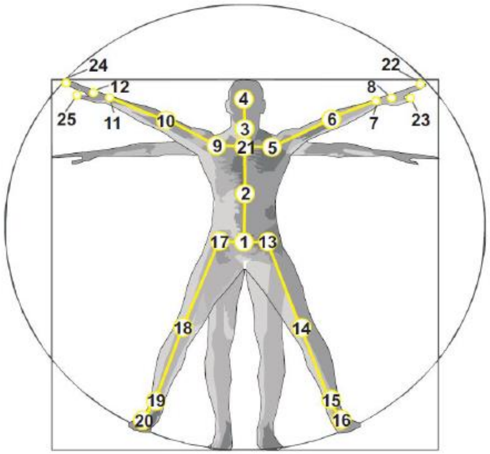


- Predicted Flow



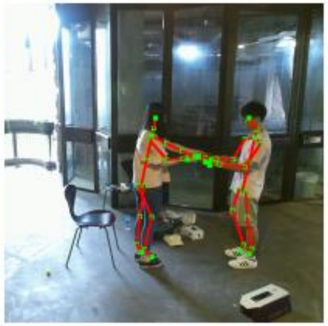
Datasets

- NTU RGB+D and NTU RGB+D 120
  - ✓ RGB videos / depth map sequences / 3D skeleton data / infrared (IR) videos 제공
  - ✓ Microsoft Kinect V2 camera를 3개 사용
  - ✓ 3D skeleton data : 25개 major body joints
  - ✓ NTU RGB+D : 60 action classes (56,880 video samples)
  - ✓ NTU RGB+D 120 : 120 action classes (114,480 video samples)
  - ✓ Cross-subject (actor 다름) 와 Cross-view (camera 위치 3개)



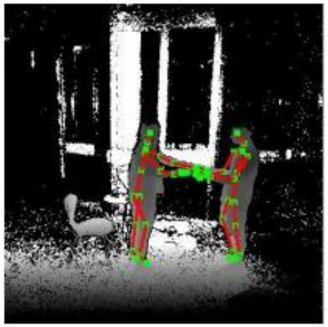
- **Sample frame의 modalities 예 :**

RGB



RGB + joints

Depth



Depth + joints

IR

Base Paper Information

- Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition
- Authors : Sijie Yan, Yuanjun Xiong, Dahua Lin

  Department of Information Engineering, The Chinese University of Hong Kong
- AAAI 2018

Abstract

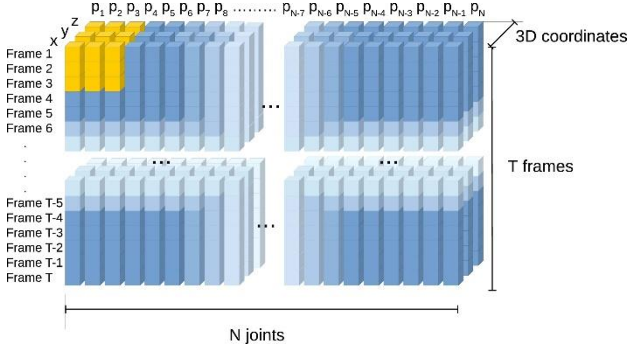
- Spatial Temporal Graph Convolutional Networks(ST-GCN)로 불리는 dynamic skeleton의 새로운 모델 제안
- Data에서 spatial pattern과 temporal pattern을 모두 학습
- Spatial-temporal graph로 구성된 block을 여러 층으로 쌓은 구조

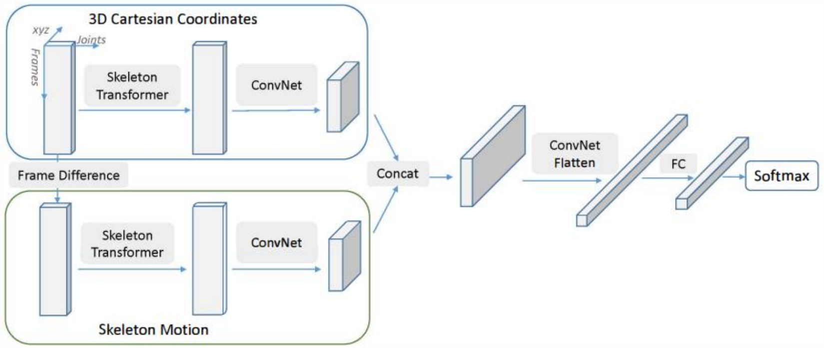
  → Spatial and temporal domain에 따라 information을 통합함
- ST-GCN은 skeleton-based action recognition task에 처음으로 GCN을 적용함

Previous Work

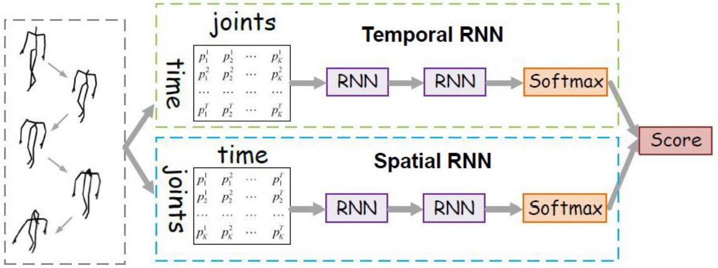
- Deep learning methods
  - CNN : model the skeleton data as a pseudo-image [1]

input data structure:





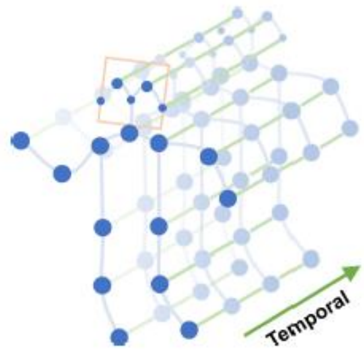
  - RNN : model the skeleton data as a sequence of the coordinate vectors along both the spatial and temporal dimensions



  - Overlook inherent connectivity correlations between joints

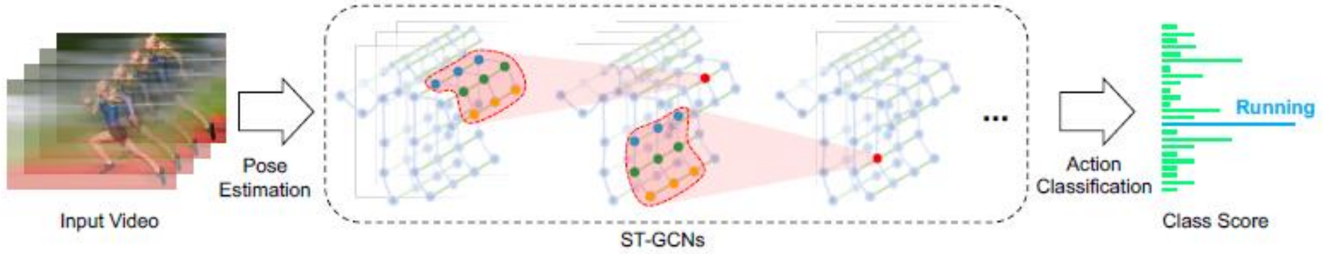
Graph-based method

- Graph-based
  - Graphs naturally captures the structure of human body
  - Joints → **nodes (or vertices)** , bones → **edges**
  - No hand-crafted node traversal



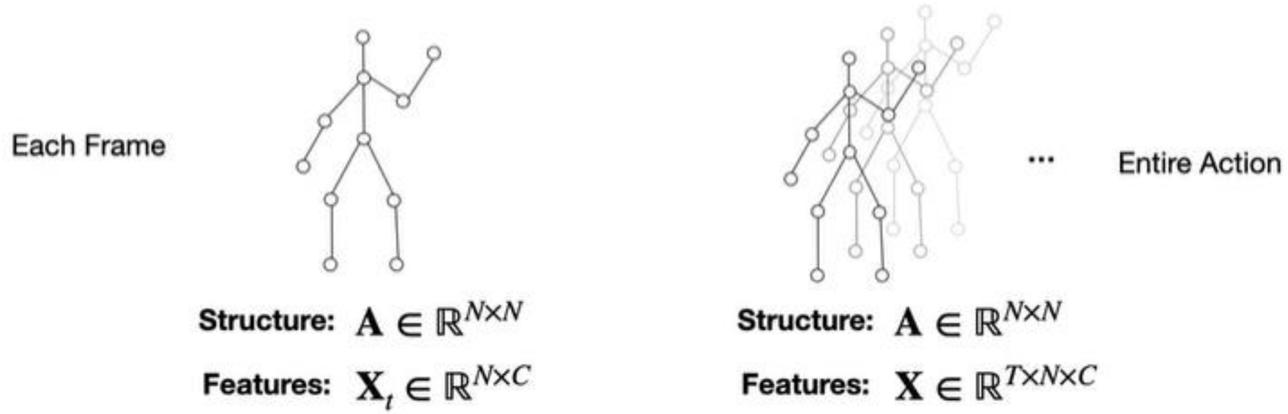
G = (V, E)

- Simple flow of ST-GCN
  - Input data가 multiple layer의 spatial-temporal graph conv. 연산을 거치면서 점차적으로 graph 상에 higher-level feature map이 생성됨
  - 이 feature map이 최종적으로 softmax를 거쳐 action class 분류



ST-GCN

- Actions as Graph sequences
  - Structure : N-node graph with adjacency matrix A (normalized $\hat{A}$)
  - Features : Joint locations X over T frames
  - Goal : Learn to classify graph sequences

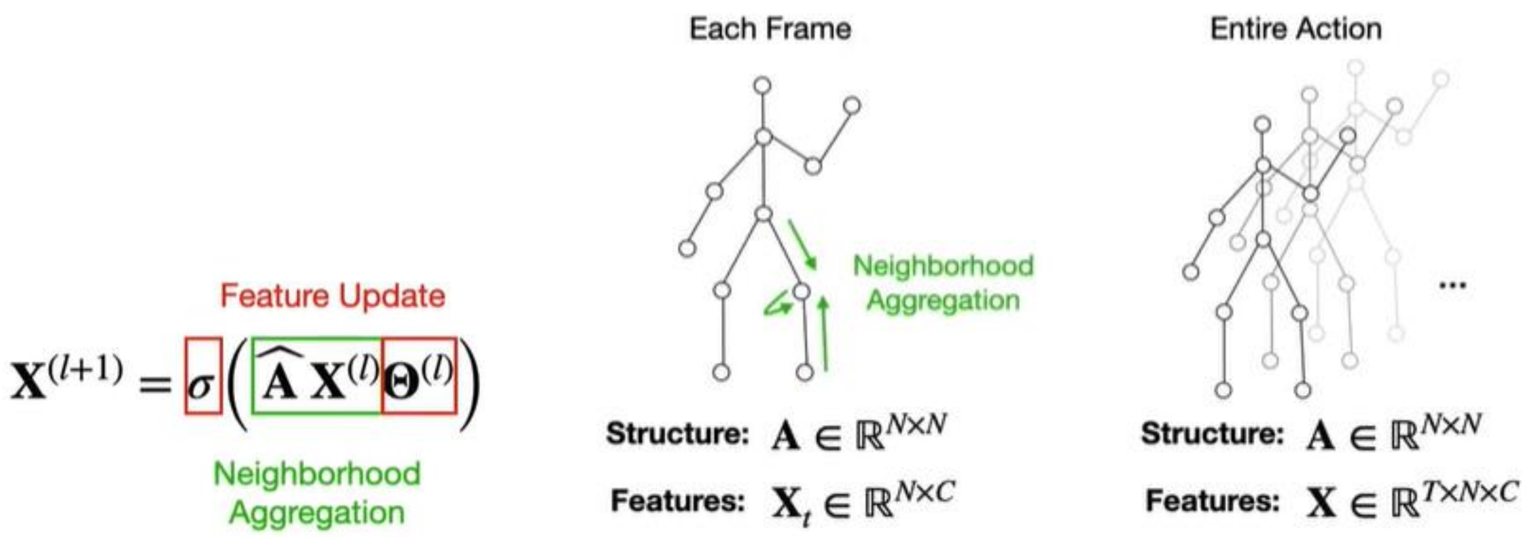


ST-GCN

- Feature learning with Graph Convolutional Nets (GCNs) ([1] Kipf *et al.*)

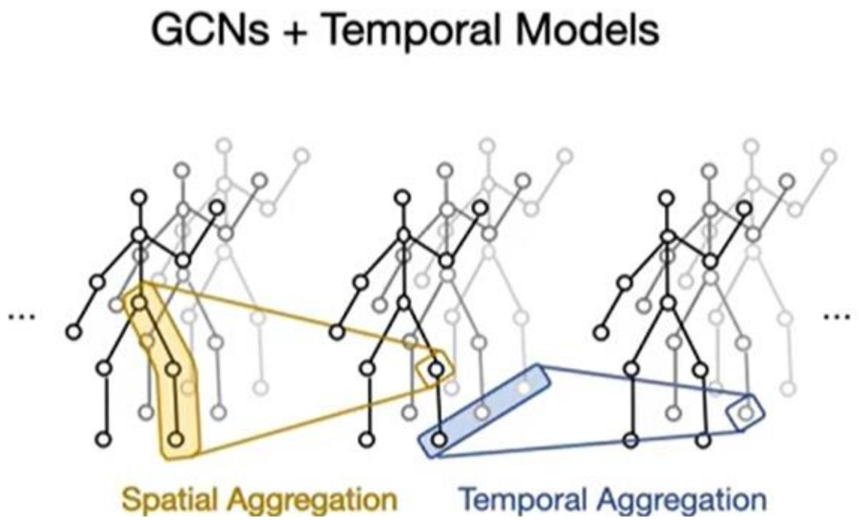
  1) Neighborhood feature aggregation

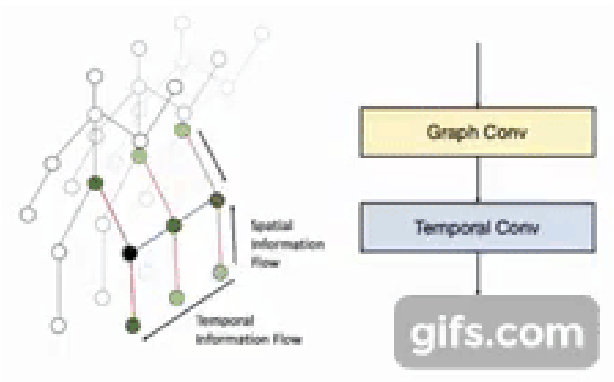
  2) Layer-wise feature update



ST-GCN

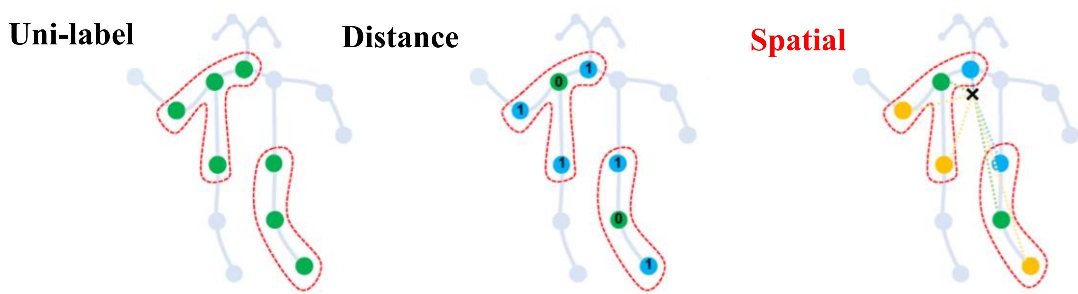
- Feature extraction
  - Learn spatial-temporal features with spatial / temporal modules
  - Spatial : Neighborhood aggregation (GCNs)
  - Temporal : Node-wise sequence models (1d Conv)





ST-GCN

- <u>Partition strategies</u> for constructing convolution operations
  - Uni-labeling : all nodes in a neighborhood has the same label (green)
  - Distance : two subsets are the root node (green) and neighboring points with distance 1 (blue)
  - Spatial configuration : distances to the skeleton gravity center (black cross), root node (green), shorter distance (blue), longer distance (yellow)



	Top-1	Top-5
Baseline TCN	20.3%	40.0%
Local Convolution	22.0%	43.2%
Uni-labeling	19.3%	37.4%
Distance partitioning*	23.9%	44.9%
Distance Partitioning	29.1%	51.3%
Spatial Configuration	29.9%	52.2%

→ 세 가지 모두 실험한 결과, 성능이 가장 좋은 **Spatial configuration** 방법을 적용함

ST-GCN

- Implementing Spatial-GCN
  - Single frame case's formula (Kipf *et al*) : $\mathbf{f}_{out} = \underbrace{\mathbf{\Lambda}^{-\frac{1}{2}}(\mathbf{A}+\mathbf{I})\mathbf{\Lambda}^{-\frac{1}{2}}}_{\text{normalized adjacency matrix}}\mathbf{f}_{in}\mathbf{W}$
  - Multiple subsets (i.e., spatial partition) : several matrixes $\mathbf{A}_j$ where $\mathbf{A}+\mathbf{I} = \sum_j \mathbf{A}_j$

$$\mathbf{f}_{out} = \sum_j \mathbf{\Lambda}_j^{-\frac{1}{2}}\mathbf{A}_j\mathbf{\Lambda}_j^{-\frac{1}{2}}\mathbf{f}_{in}\mathbf{W}_j,$$

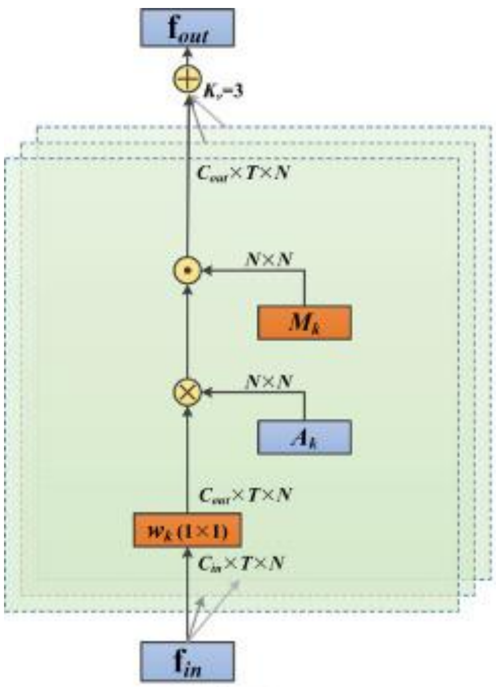


Fig. Spatial graph convolution

ST-GCN

- Network architecture 및 algorithm 순서
    1) 동영상으로부터 skeleton을 추출
    2) skeleton data를 그래프 형태로 만듬
       : 각 joints를 nodes로 만들고 nodes가 이어지는 부분(공간, 시간)을 edge로 연결
    3) 총 10개의 ST-GCN block을 통해 feature를 추출
    4) Softmax 함수를 이용하여 행동을 분류

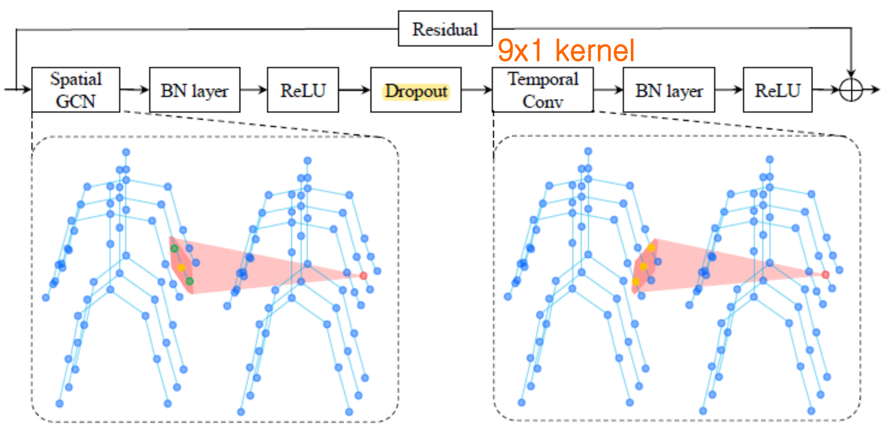


Fig 1. ST-GCN module

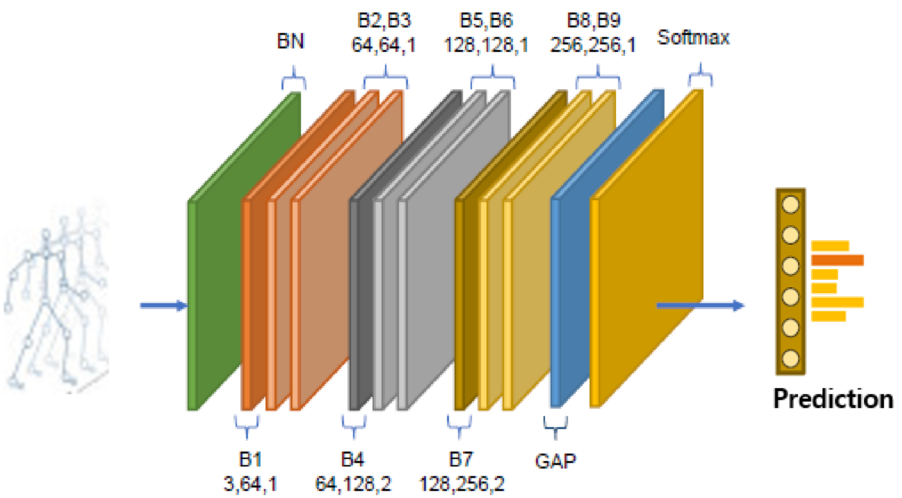


Fig 2. Illustration of the network

Experimental result 및 단점

- NTU RGB + D dataset
  - Conventional handcrafted method를 사용한 방법이나, RNN 또는 CNN based methods에 비해 가장 좋은 성능을 보여줌. (2018년 기준)

	actor X-Sub	camera X-View
Lie Group (Veeriah, Zhuang, and Qi 2015)	50.1%	52.8%
H-RNN (Du, Wang, and Wang 2015)	59.1%	64.0%
Deep LSTM (Shahroudy et al. 2016)	60.7%	67.3%
PA-LSTM (Shahroudy et al. 2016)	62.9%	70.3%
ST-LSTM+TS (Liu et al. 2016)	69.2%	77.7%
Temporal Conv (Kim and Reiter 2017).	74.3%	83.1%
C-CNN + MTLN (Ke et al. 2017)	79.6%	84.8%
ST-GCN	**81.5%**	**88.3%**

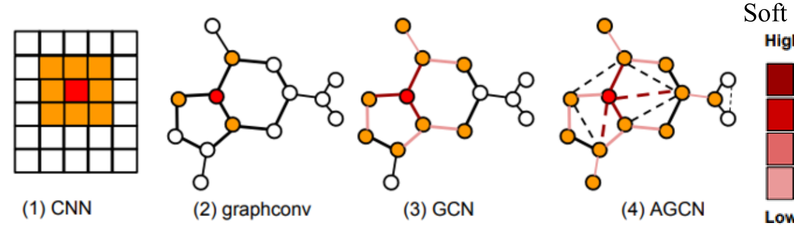
- ST-GCN 모델의 단점
  - 관절의 관계성을 로컬 영역(연결된 관절 간의 관계)에서 밖에 찾지 못함
    Ex) 멀리 떨어진 왼손과 오른발의 관계성을 찾는데 명시적이지 않음

ST-GCN 이후 Research

- Spatial GCN 성능 개선
- Temporal Conv. 성능 개선
- Attention mechanism (key feature attention)
- Architecture 변경
- Multi-stream 적용

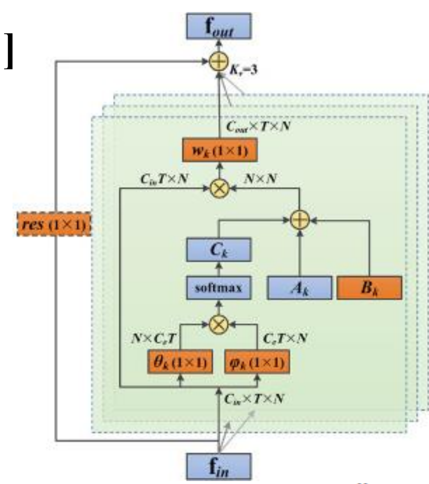
[1] Lei *et al*. "Two-stream adaptive graph convolutional networks for skeleton-based action recognition" CVPR 2019
[2] Zhu et al. "Topology-learnable graph convolution for skeleton-based action recognition" Pat. Recog. Letters 2020
[3] Xu *et al*. "Multi-scale skeleton adaptive weighted GCN for skeleton-based human action recognition in IoT" Applied Soft Computing Journal, 2021

Spatial GCN 개선



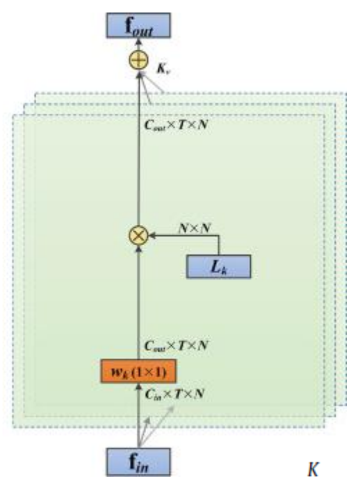
- Spatial graph convolution
  - Adaptive GCN [1] : 기본 Adj. Mat.(A_k) + 연결성 및 연결 강도를 학습(B_k) + data (각 sample) dependent graph로 joints의 similarity(C_k)학습 (1x1 conv. 임베딩 함수 이용)
  - Topology learnable GCN [2] : Non-local mechanism이 추가로 필요하지 않다.
    L_k를 A_k로 초기화한 후 학습하면 성능이 더 좋다
  - Adaptive weighted GCN [3] : 위와 비슷하지만 성능이 더 좋다 (Ablation 실험 없음)

[1]



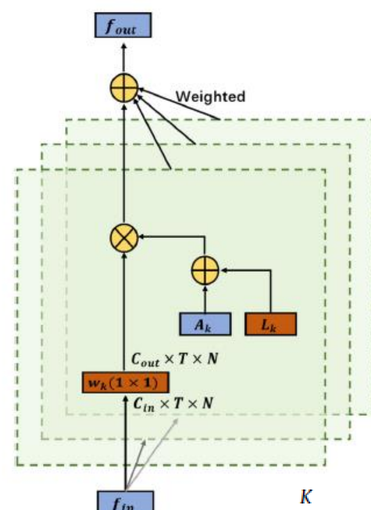
$$f_{out} = \sum_{k}^{K} W_k f_{in}(A_k + B_k + C_k)$$

[2]



$$f_{out} = \sum_{k}^{K} W_k f_{in}(L_k)$$

[3]



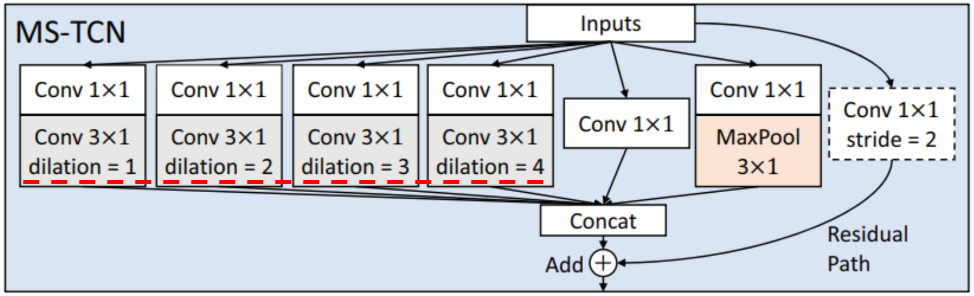
$$f_{out} = \sum_{k}^{K} \alpha W_k f_{in}(A_k + L_k)$$

Temporal features 개선

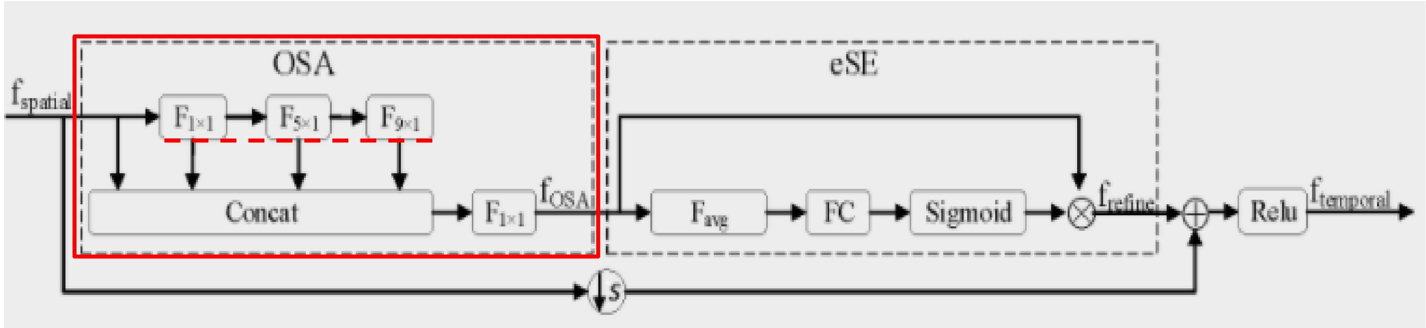
- Temporal convolution 구조 변경
  - Multi-scale 구조 [1] : bottleneck design (computational cost ↓) + kernel 3x1 + dilation rate 변경
  - OSA (one-shot aggregation) [2] : multiple kernels with different smaller size

Best kernel 조합은 1x1, 5x1, 9x1을

[1]



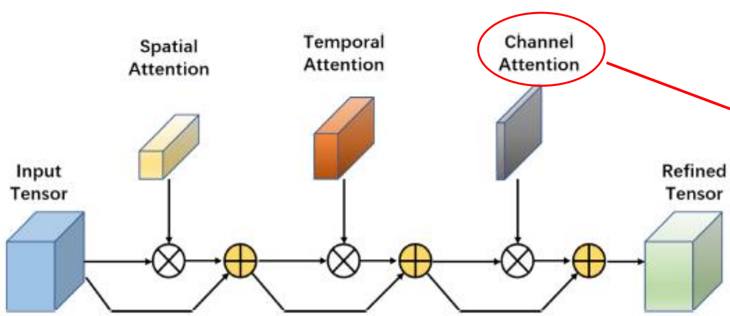
[2]

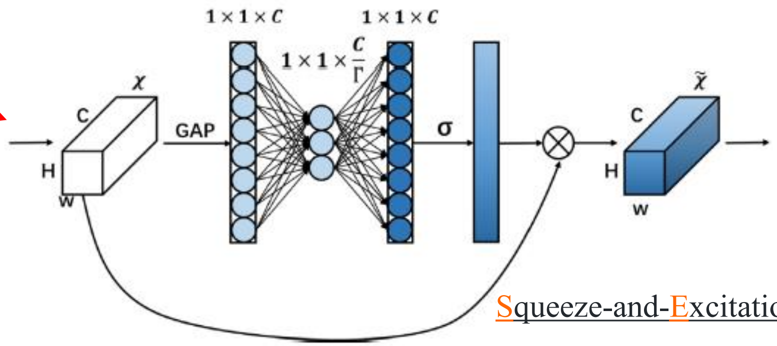


Attention mechanism

- Attention helps the model pay more attention to important joints, frames and features
  - STC attention [1] : SAM + TAM + CAM 을 차례로 적용
  - ECA attention [2] 적용 : SAM, TAM light weight CAM인 ECA-Net [3] 적용

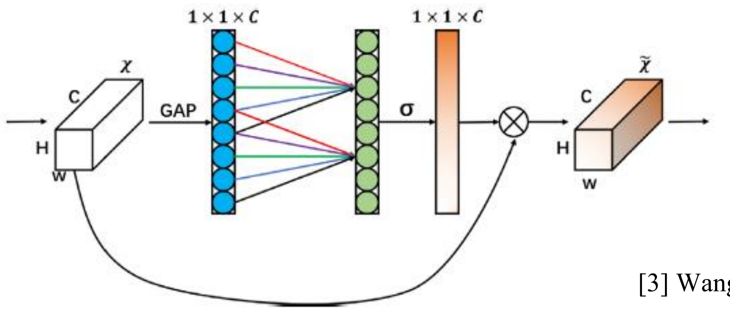
[1]





Squeeze-and-Excitation

[2]



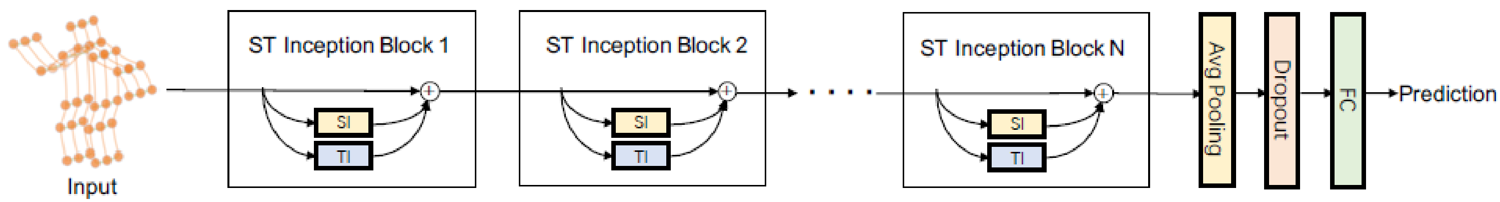
Efficient Channel Attention

[3] Wang *et al*. "ECA-Net: Efficient channel attention for deep CNNs" CVPR 2020

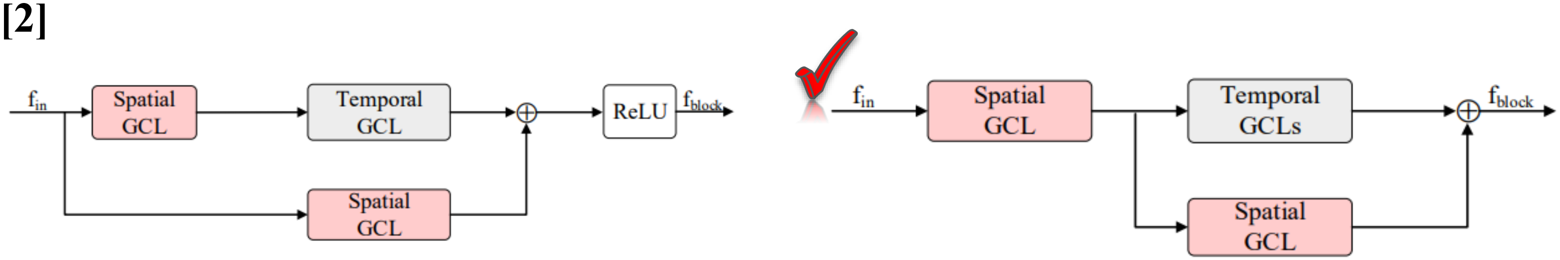
Architecture change

- 기존 CNN architecture들을 GCN 구조에 활용
  - Inception block (GoogLeNet) 구조를 참조한 architecture [1] : S=4, T=2 (branch 수)
  - Various variants 검토 [2] : extended spatial-temporal 구조

[1]



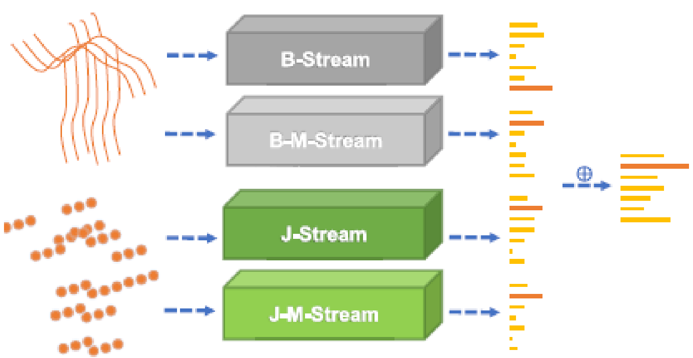
[2]



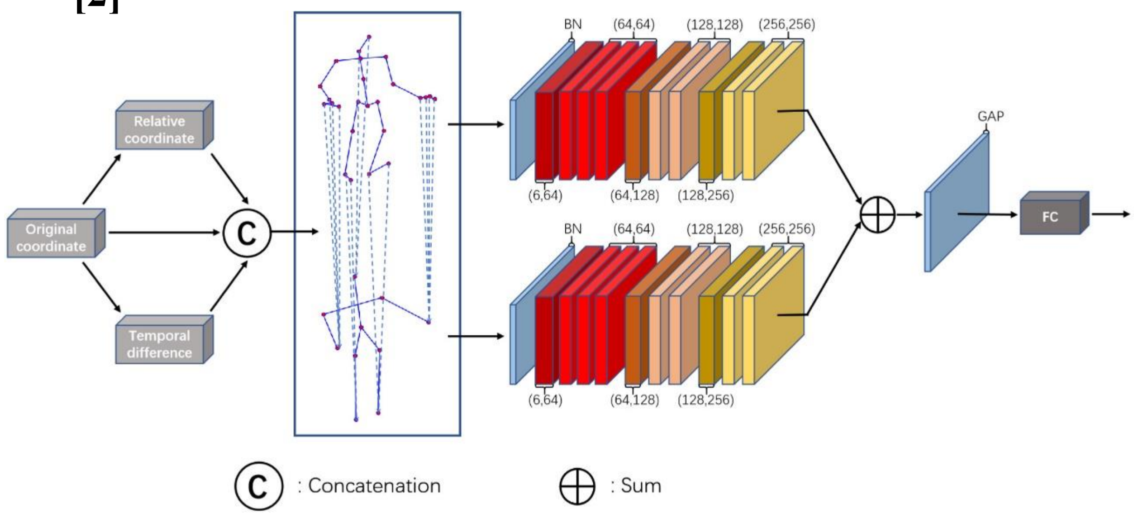
Multi-stream network

- Joints 좌표로 부족한 정보를 여러 information을 활용하여 성능 향상
  - 4-stream network fusion [1] : joint / bone / joint motion / bone motion 정보 활용
    학습된 4개 network를 ensemble하여 class prediction
  - Multi-scale network [2] : 더 풍부한 spatial feature를 추출하기 위해 multi scale 적용
    joints / relative joints / j_motion을 concat 후 9ch data input
    25 joints와 10 joints (작은 joints 무시) 사용 후 feature fusion

[1]



[2]



Experiments

- Base ST-GCN 모델 대비 GCN 구조 변경, attention 적용, multi-stream, multi-scale network 등을 적용하여 성능이 많이 향상됨

[1]

Methods	CS (%)	CV (%)
Lie Group [6]	50.1	82.8
HBRNN [7]	59.1	64.0
Deep LSTM [14]	60.7	67.3
ST-LSTM [49]	69.2	77.7
STA-LSTM [8]	73.4	81.2
VA-LSTM [18]	79.2	87.7
Ind-RNN [9]	81.8	88.0
SRN+TSL [19]	84.8	92.4
TCN [20]	74.3	83.1
Clips+CNN+MTLN [50]	79.6	84.8
Synthesized CNN [21]	80.0	87.2
CNN+Motion+Trans [22]	83.2	89.3
3scale ResNet152 [23]	85.0	92.3
ST-GCN [10]	81.5	88.3
DPRL+GCNN [25]	83.5	89.8
ASGCN [27]	86.8	94.2
AGCN [16]	88.5	95.1
AGC-LSTM [26]	89.2	95.0
MS-AAGCN (ours)	90.0± 0.109	96.2± 0.095

[2]

Comparisons on NTU-RGB+D 60.

Methods	Cross-subject (%)	Cross-view (%)
HBRNN [80]	59.1	64.0
Deep LSTM [77]]	60.7	67.3
ST-LSTM [81]	69.2	77.7
STA-LSTM [82]	73.4	81.2
SRN-TSL [40]	84.8	92.4
ST-GCN [48]	81.5	88.3
2S-AGCN [52]	88.5	95.1
2S-AGCN [52]+Attention	89.4	96.0
MS-AWGCN(ours)	90.3	96.4

Conclusion

- Skeleton-based human action recognition task 에서…
- GCN을 적용한 network 적용으로 성능 향상이 많이 되었음
- ST-GCN의 경우 GCN을 통해 spatial feature 를 extraction 함
- ST-GCN 이후에도 spatial-temporal module 구조 변경 및 attention mechanism, architecture 변경, Multi-stream network 구성 등을 통해 성능이 향상되었음

- Action recognition 관련 참조 site
  - https://niais.github.io/Awesome-Skeleton-based-Action-Recognition/

Thank You