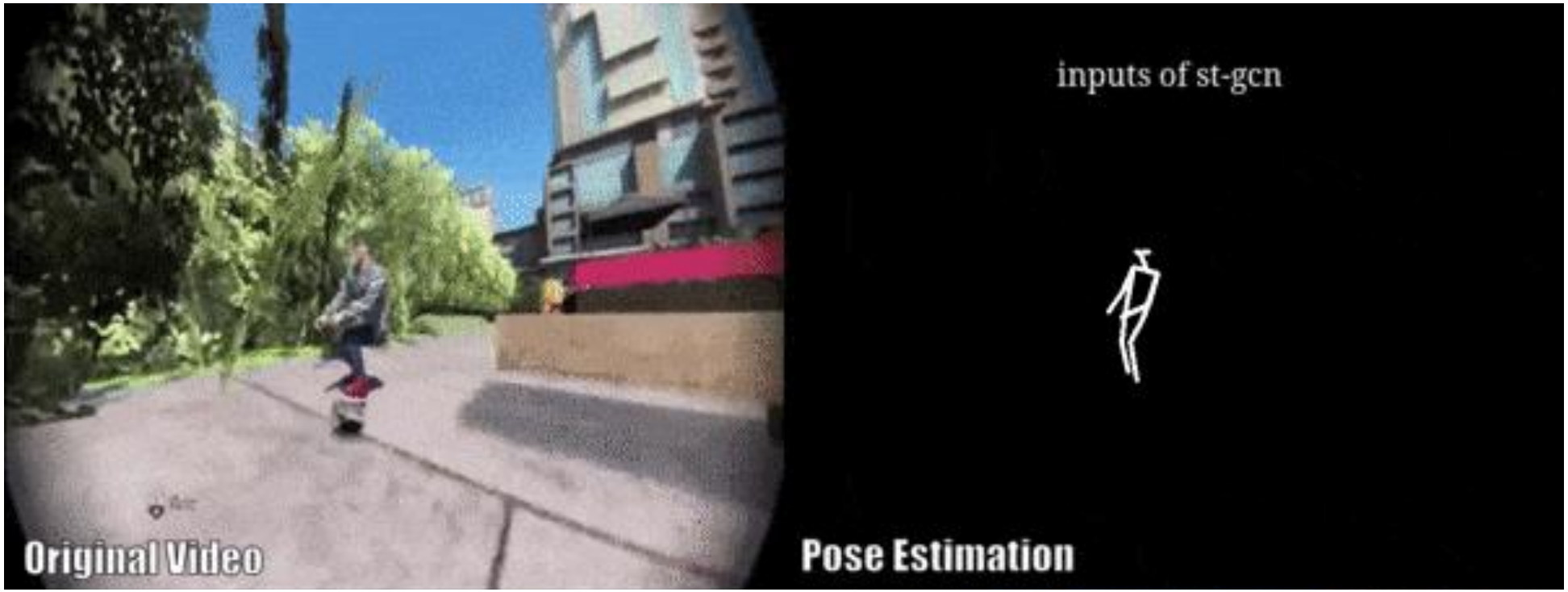
Disentangling and Unifying Graph Convolutions for Skeleton-Based Action Recognition

Dong-Hwan Jang

Seoul National University

Introduction

Skeleton-Based Action Recognition



- Predicting actions from skeleton representations of human bodies instead of raw RGB videos
  - In recent work, the significant results have proven its merits

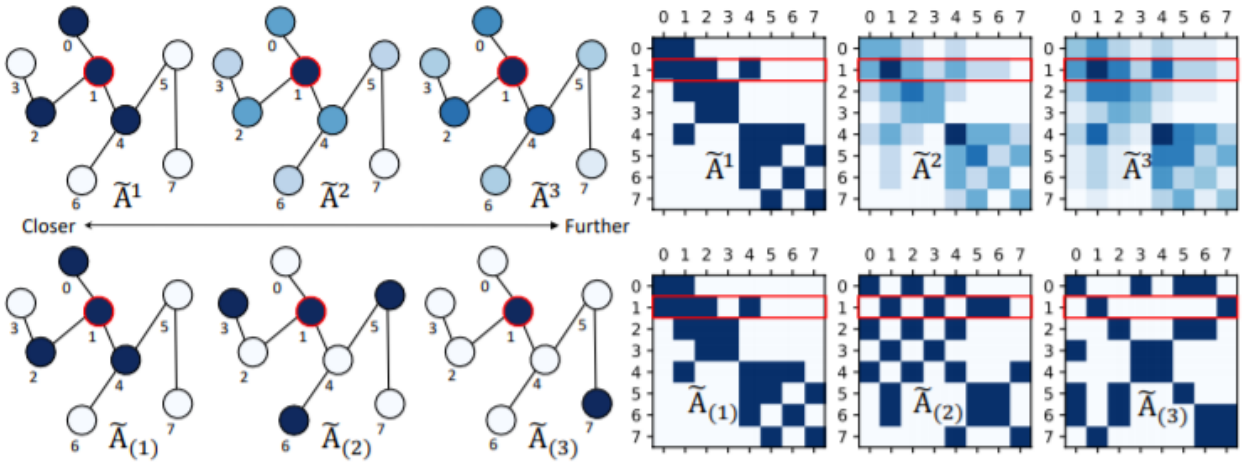
Introduction

Conditions for robust action recognition from skeleton

1) Extract **multi-scale structural features & long-range dependencies**

- Joints that are structurally apart can also have strong correlations
- Many existing approaches achieve this with higher-order polynomials of the adjacency matrix, A
  - **Biased weight problem**

Disentangling local dependencies

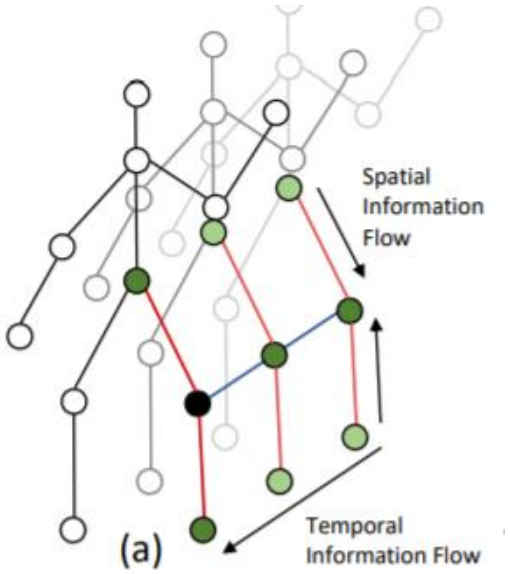


Introduction

Conditions for robust action recognition from skeleton

2) Leverage the complex **cross-spacetime joint relationships**

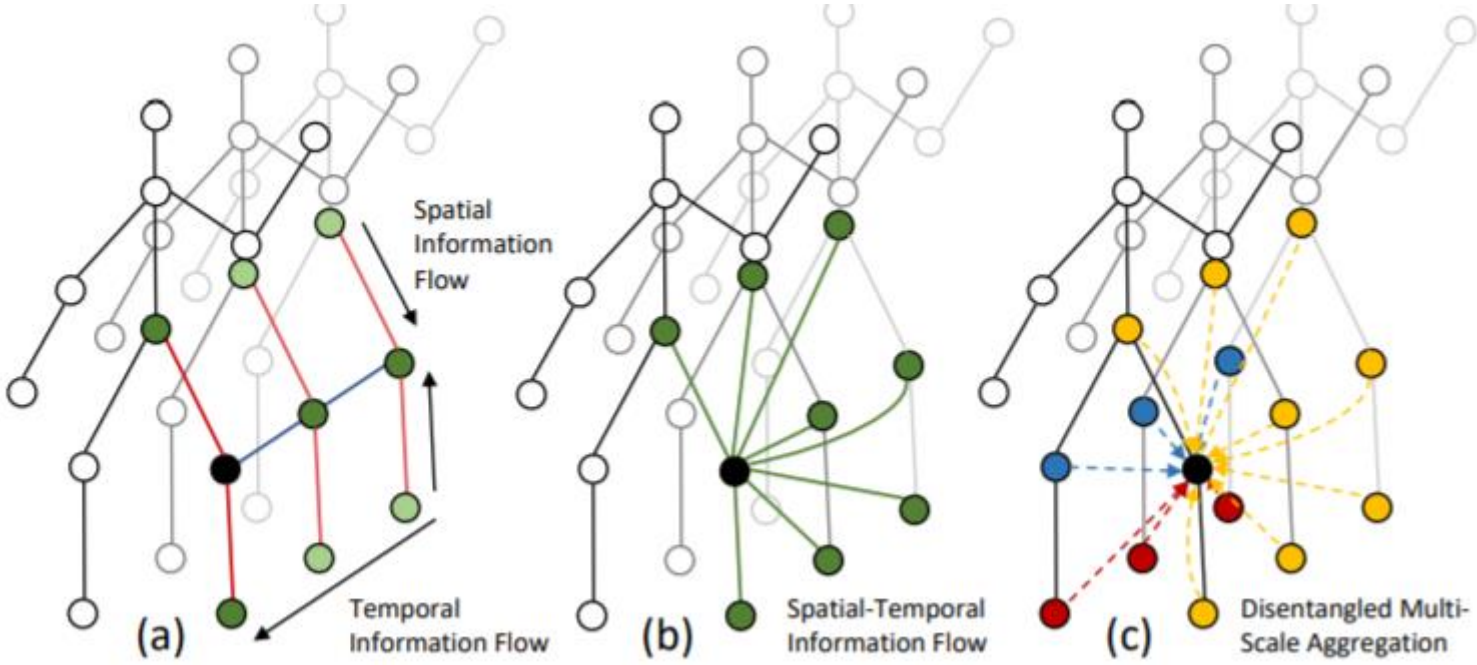
- Most existing approaches deploy interleaving spatial-only & temporal-only modules (similar to factorized C3D)
  - GCN – RNN or Conv1D
- E.g. “Standing Up” – past (*Upper body* “leaning forward) + future (*Lower body* “standing up”)



Skip Connection

Method

Multi-Scale G3D: MS-G3D

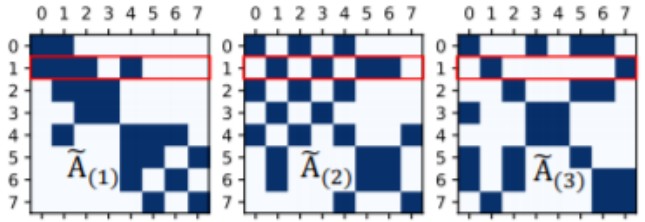


2) cross-spacetime 1) long-range dep.

Method

Multi-Scale G3D: MS-G3D

$$[\mathbf{X}_{(\tau)}^{(l+1)}]_t = \sigma\left(\sum_{k=0}^{K} \tilde{\mathbf{D}}_{(\tau,k)}^{-\frac{1}{2}} \tilde{\mathbf{A}}_{(\tau,k)} \tilde{\mathbf{D}}_{(\tau,k)}^{-\frac{1}{2}} [\mathbf{X}_{(\tau)}^{(l)}]_t \Theta_{(k)}^{(l)}\right)$$

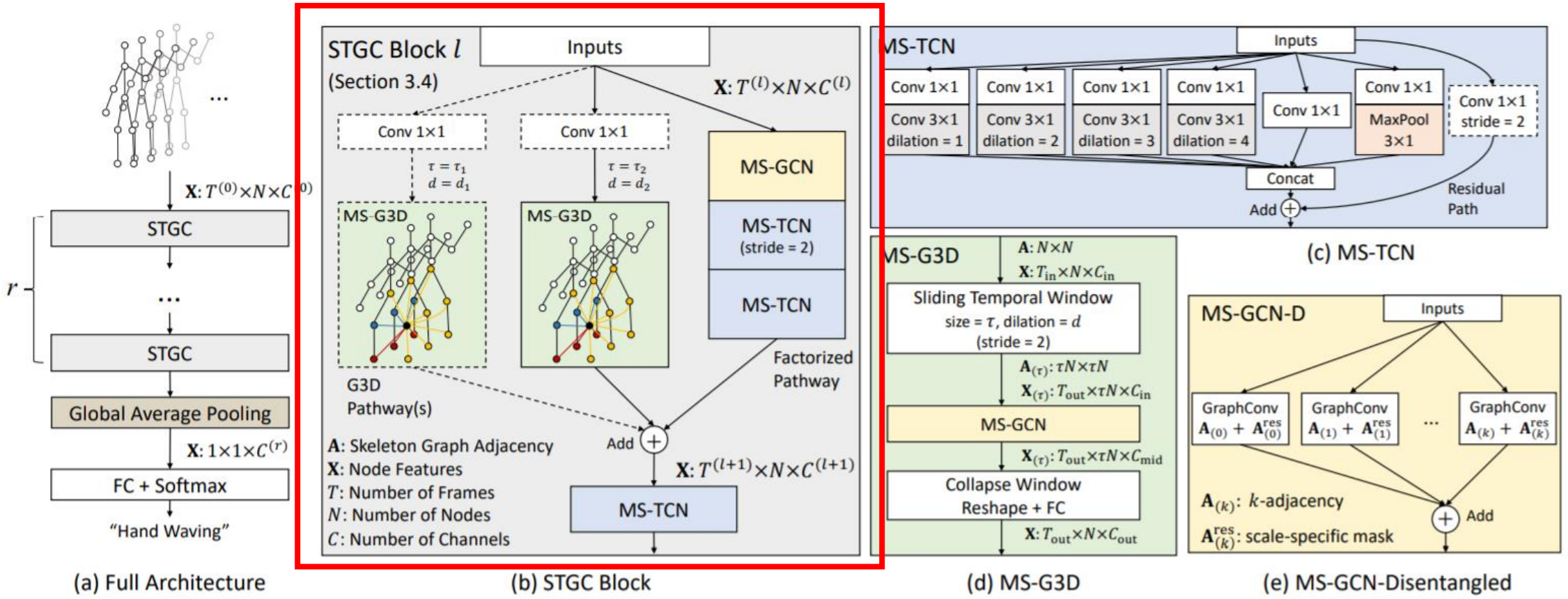


1) $\tilde{\mathbf{A}}_{(\tau,k)}$ is disentangled

2) Skip connections are considered in $\tilde{\mathbf{A}}_{(\tau,k)}$ $\tilde{\mathbf{D}}_{(\tau,k)}^{-\frac{1}{2}}$

Method

Multi-Scale G3D: MS-G3D



Results

Methods	Number of Scales			
	$K = 1$	$K = 4$	$K = 8$	$K = 12$
GCN-E	85.1	85.6	86.5	86.6
GCN-D	85.1	87.0	86.9	86.8
GCN-E + Mask	86.1	87.0	87.5	87.7
GCN-D + Mask	86.1	86.9	87.9	87.8
G3D-E	85.1	85.5	85.4	85.5
G3D-D	85.1	86.4	86.5	86.4
G3D-E + Mask	86.6	87.0	86.5	86.2
G3D-D + Mask	86.6	87.4	87.1	87.0

Table 1: Accuracy (%) with multi-scale aggregation on individual pathways of STGC blocks with different K. "Mask" refers to the residual masks $\mathbf{A}^{\text{res}}$. If $K>1$, GCN/G3D is **Multi-Scale** (**MS-**).

Results

Model Configurations	Params	Acc (%)
Baseline (Js-AGCN [33])	3.5M	86.0
Baseline + MS-TCN	1.6M	86.7
MS-GCN (Factorized Pathway) Only	1.4M	87.8
with 2.5× Capacity	3.5M	88.5
with Dual Pathway	2.8M	88.6
MS-GCN (Factorized Pathway)		
with MS-G3D ($\tau = 3, d = 1$)	2.7M	89.0
with MS-G3D ($\tau = 3, d = 2$)	2.7M	89.1
with MS-G3D ($\tau = 3, d = 3$)	2.7M	89.1
with MS-G3D ($\tau = 5, d = 1$)	3.2M	89.2
with MS-G3D ($\tau = 5, d = 2$)	3.2M	89.2
with MS-G3D ($\tau = 7, d = 1$)†	3.0M	89.0
with 2 MS-G3D Pathways† $\tau = (3, 3), d = (1, 2)$	2.8M	89.3
with 2 MS-G3D Pathways† $\tau = (3, 5), d = (1, 1)$	3.2M	89.4

Results

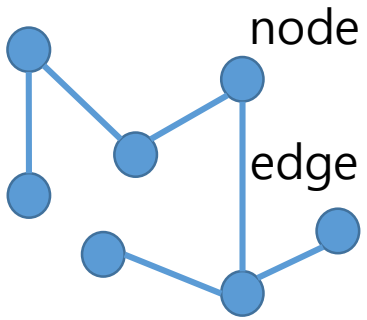
Methods	NTU RGB+D 120	
	X-Sub (%)	X-Set (%)
ST-LSTM [26]	55.7	57.9
GCA-LSTM [27]	61.2	63.3
RotClips + MTCNN [16]	62.2	61.8
Body Pose Evolution Map [28]	64.6	66.9
2s-AGCN [33]	82.9	84.9
MS-G3D Net	**86.9**	**88.4**

Methods	NTU RGB+D 60	
	X-Sub (%)	X-View (%)
IndRNN [23]	81.8	88.0
HCN [20]	86.5	91.1
ST-GR [18]	86.9	92.3
AS-GCN [21]	86.8	94.2
2s-AGCN [33]	88.5	95.1
AGC-LSTM [34]	89.2	95.0
DGNN [32]	89.9	96.1
GR-GCN [8]	87.5	94.3
MS-G3D Net (Joint Only)	89.4	95.0
MS-G3D Net (Bone Only)	90.1	95.3
MS-G3D Net	**91.5**	**96.2**

Methods	Kinetics Skeleton 400	
	Top-1 (%)	Top-5 (%)
ST-GCN [50]	30.7	52.8
AS-GCN [21]	34.8	56.5
ST-GR [18]	33.6	56.1
2s-AGCN [33]	36.1	58.7
DGNN [32]	36.9	59.6
MS-G3D Net	**38.0**	**60.9**

Graphs from social networks

- people and their interactions
- directed (Twitter) and undirected (Facebook)
- typical ML tasks
  - Link(edge) prediction
  - advertising (recommendation)
  - product placement



Social Graphs