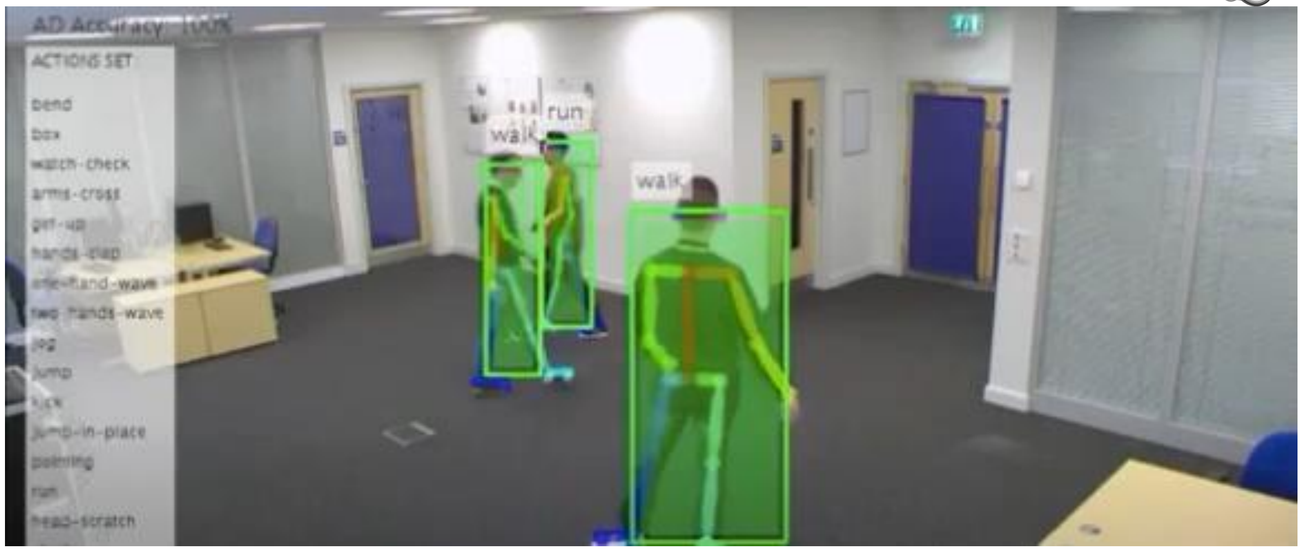
Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition

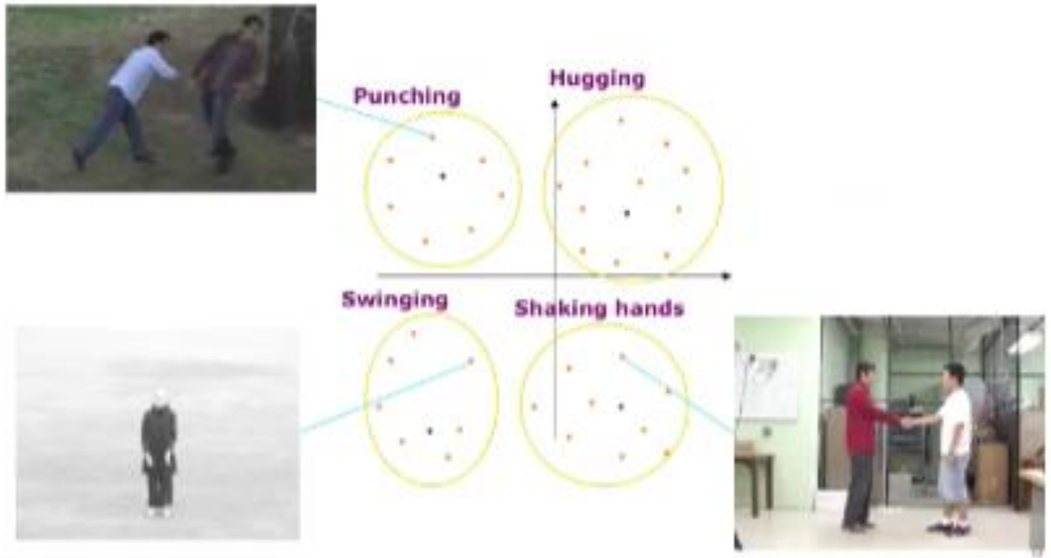
Sijie Yan, Yuanjun Xiong, Dahua Lin
Department of Information Engineering, The Chinese University of Hong Kong
{ys016, dhlin}@ie.cuhk.edu.hk, bitxiong@gmail.com

GCN lecture 중간 발표
2020.05.08
2018-23112 홍석일

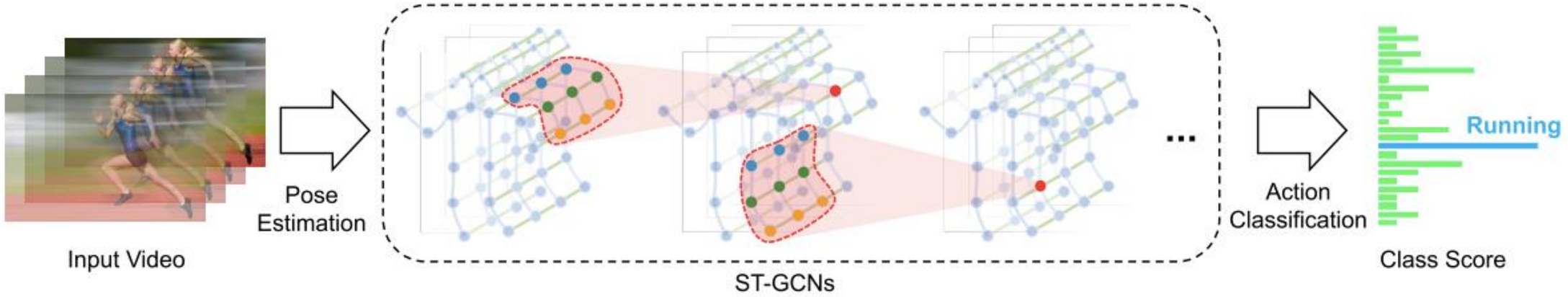
Human Action Recognition

- Input = A (human) video segment containing activities
- Problems : Action classification, Action detection/localization, Action prediction/forecast
- Dataset : UCF101, HMDB, Kinetics, …
- Multiple modalities : Appearance, Depth, Optical flows, and Body skeletons





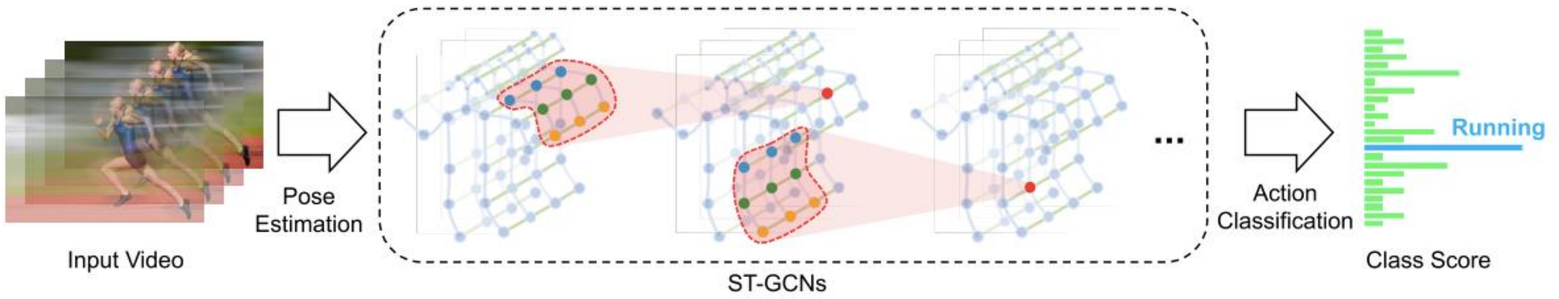
Contributions



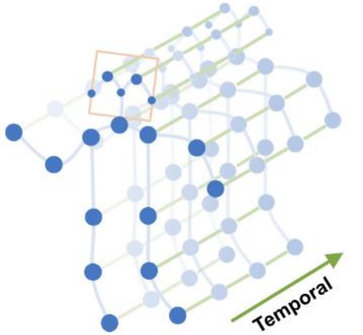
- Proposed ST-GCN, a generic graph-based formulation for modeling dynamic skeletons, which is the first that applies graph-based neural networks for this task.
- Proposed several principles in designing convolution kernels in ST-GCN to meet the specific demands in skeleton modeling.
- On two large scale datasets for skeleton-based action recognition, the proposed model achieves superior performance as compared to previous methods.

ST-GCN (Spatial-Temporal Graph Convolutional Networks)

- Pipeline



- Skeleton Graph Construction



  - spatial temporal graph of a skeleton sequence

- Spatial Graph Convolutional Neural Network

$$f_{out}(\mathbf{x}) = \sum_{h=1}^{K}\sum_{w=1}^{K} f_{in}(\mathbf{p}(\mathbf{x},h,w)) \cdot \mathbf{w}(h,w)$$

  - CNN

$$f_{out}(v_{ti}) = \sum_{v_{tj}\in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(\mathbf{p}(v_{ti},v_{tj})) \cdot \mathbf{w}(v_{ti},v_{tj})$$

  - Graph Convolution

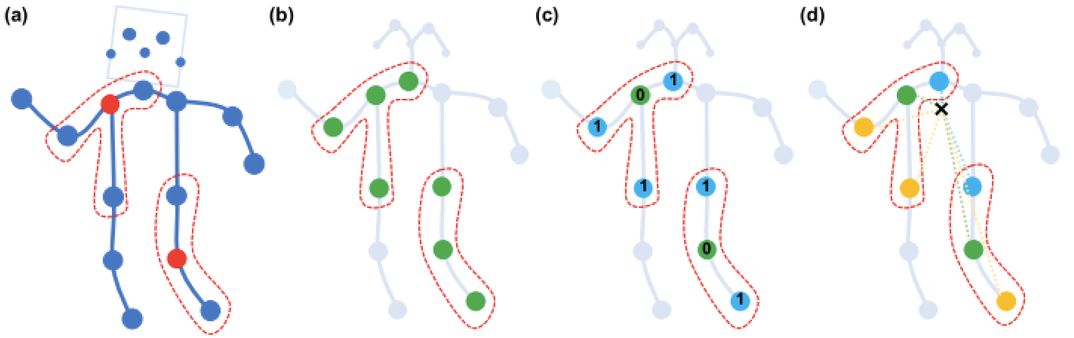
$$f_{out}(v_{ti}) = \sum_{v_{tj}\in B(v_{ti})} \frac{1}{Z_{ti}(v_{tj})} f_{in}(v_{tj}) \cdot \mathbf{w}(l_{ti}(v_{tj})).$$

  - Spatial Graph Convolution

$$B(v_{ti}) = \{v_{qj} | d(v_{tj}, v_{ti}) \leq K, |q-t| \leq \lfloor \Gamma/2 \rfloor\}.$$

  - Spatial Temporal Modeling

- Partition Strategies



  - Uni-labeling partitioning
  - Distance partitioning
  - Spatial configuration partitioning

$$l_{ti}(v_tj) = \begin{cases} 0 & \text{if } r_j = r_i \\ 1 & \text{if } r_j < r_i \\ 2 & \text{if } r_j > r_i \end{cases}$$

Experiments

- Skeleton based action recognition performance on NTU-RGB+D datasets

	X-Sub	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%**

- Mean class accuracies on the “Kinetics Motion” subset of the Kinetics dataset.

Method	RGB CNN	Flow CNN	ST-GCN
Accuracy	70.4%	72.8%	72.4%

- Exploration using ST-GCN to capture motion information in two-stream style action recognition

	RGB TSN	Flow TSN	ST-GCN	Acc(%)
Single Model	✓			70.3
		✓		51.0
			✓	30.7
Ensemble Model	✓	✓		71.1
	✓		✓	71.2
	✓	✓	✓	71.7


Thank
you!