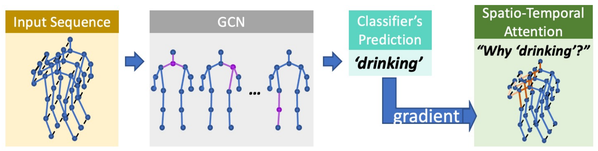
Towards To-a-T Spatio-Temporal Focus for Skeleton-Based Action Recognition

Lipeng Ke[1], Kuan-Chuan Peng[2], Siwei Lyu[1]
[1]SUNY at Buffalo, [2]Mitsubishi Electric Research Laboratories
lipengke@buffalo.edu, kpeng@merl.com, siweilyu@buffalo.edu

Introduction



Adjacency matrix (Spatial): *used to model joints' dependency*
- Do not have learnable dynamic topology of joint connection.
- Do not have learnable dynamic intensity of joint connection.
- Do not embed spatio-temporal focus with direct enforcement.

Attention module (Temporal): *used to tell which joints in what movement is important*
- Do not have objectives to directly enforce the net to capture the spatial and temporal patterns jointly.
- Do not have objectives to learn discriminative joints and joint movement for different action classes.

STF Exploration L_e

When and where to look at

$$L_e = g^y(X - Q \odot X)$$

y is the prediction class of the original sequence, $g^y(\cdot)$ extracts the prediction score of $(X - Q \odot X)$ at the original prediction class y, and X is the input sequence, $\odot$ is the element-wise production.

STF Consistent L_c

- Same input same attention on different GCNs
- Later GCN larger receptive field and better attention

$$L_c = \left\| Q^i - Q^j \right\|_2$$

Q^i, Q^j are the spatio-temporal focus of the STF-GCN modules i, j respectively

STF Divergence L_d

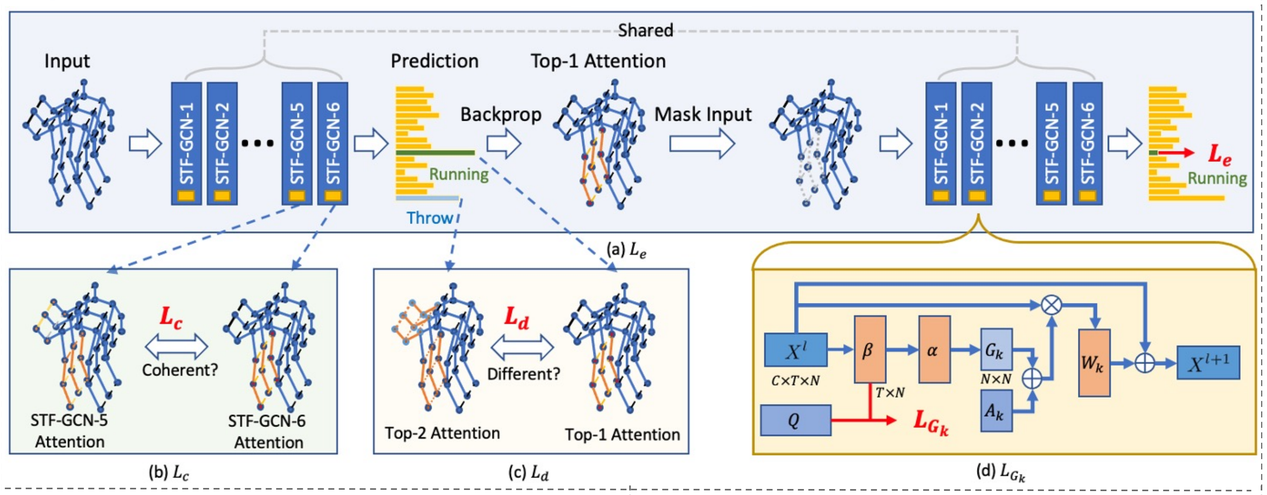
- Different class different focus

classes	Spatial similarity	Temporal similarity
walking vs. running	similar	different
reading vs. walking	different	different

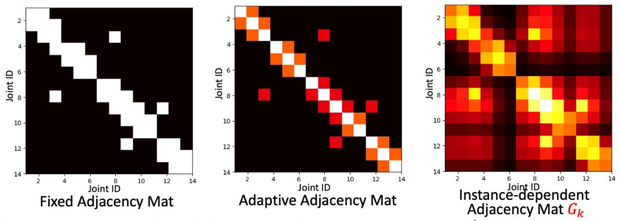
$$L_d = -\|Q^{y_i} - Q^{y_j}\|_2$$

Q^{y_i}, Q^{y_j} are the spatio-temporal focus of the top-2 prediction classes y_i, y_j respectively

Pipeline



STF Module (STF-GCN)



Instance dependent Adjacency Matrix: $G_k = \alpha(\beta(X^l))$

STF loss: $L_{G_k} = \|Q - \beta(X^l)\|_2$

Overall STF Loss

$$L = L_{ce} + \lambda_e L_e + \lambda_d L_d + \lambda_c L_c + \lambda_{G_k} L_{G_k}$$

where $\lambda_e, \lambda_d, \lambda_c, \lambda_{G_k}$ are the weights of the losses.
- Adjust λ such each loss term is of same order
- Optimizing L_e and L_d together makes the training process unstable: separate L_e from the other losses

Dataset

- 3D skeleton: NTU RGB+D 60/120
- 2D skeleton: Kinetics Skeleton 400

property \ dataset	NTU RGB+D 60		NTU RGB+D 120		Kinetics Skeleton 400
settings	x-sub	x-view	x-set	x-sub	-
# class	60	60	120	120	400
# training/testing sequences	40091/16487	37646/18932	54468/59477	63026/50919	240436/19796
# training/testing subjects/views/setups	20/20 (subjects)	2/1 (views)	16/16 (setups)	53/53 (subjects)	-
# keypoints	25	25	25	25	18
spatial coordinate	3D	3D	3D	3D	2D

Experiment: Comparison w/ SOTA

dataset	NTU RGB+D 60						NTU RGB+D 120						Kinetics-400		
setting	x-sub			x-view			x-sub			x-set					
method \ input modality	J	B	$J+B$	J	B	$J+B$	J	B	$J+B$	J	B	$J+B$	J	B	$J+B$
SR-TSL (Si et al. 2018) ECCV'18	84.80	-	-	92.40	-	-	-	-	-	-	-	-	-	-	-
2s-AGCN (Shi et al. 2019) CVPR'19	-	-	88.50	93.70	93.20	95.10	-	-	-	-	-	-	35.10	33.30	36.10
TS-SAN (Cho et al. 2020) WACV'20	87.20	-	-	92.70	-	-	-	-	-	-	-	-	35.10	-	-
GCN-NAS (Peng et al. 2020) AAAI'20	-	-	89.40	94.60	94.70	95.70	-	-	-	-	-	-	35.50	34.90	37.10
MS-TGN (Li, Zhang, and Li 2020) ICCSIT'20	86.60	87.50	89.50	94.10	93.90	95.90	-	-	-	-	-	-	35.20	33.30	37.30
MS-AAGCN (Shi et al. 2020b) TIP'20	88.00	88.40	89.40	95.10	94.70	96.00	-	-	-	-	-	-	36.00	34.70	37.40
3s RA-GCN (Song et al. 2020a) TCSVT'20	-	-	87.30⋆	-	-	93.60⋆	-	-	81.10⋆	-	-	82.70⋆	-	-	-
DC-GCN+ADG (Cheng et al. 2020) ECCV'20	-	-	90.80‡	-	-	96.60‡	-	-	86.50‡	-	-	88.10‡	-	-	-
DSTA-Net (Shi et al. 2020a) ACCV'20	-	-	91.50◇	-	-	96.40◇	-	-	86.60◇	-	-	89.00◇	-	-	-
STIGCN (Huang et al. 2020) ACMMM'20	90.10	-	-	96.10	-	-	-	-	-	-	-	-	37.90	-	-
PA-ResGCN-B19 (Song et al. 2020b) ACMMM'20	-	-	90.90†	-	-	96.00†	-	-	87.30†	-	-	88.30†	-	-	-
PST-GCN (Heidari and Iosifidis 2020b) arXiv'20	87.90	-	88.68	94.33	-	95.10	-	-	-	-	-	-	34.71	-	35.53
Dynamic-GCN (Ye et al. 2020) ACMMM'20	-	-	91.50	-	-	96.00	-	-	87.30	-	-	88.60	-	-	37.90
MS TE-GCN (Li et al. 2020) arXiv'20	87.40	88.50	90.80‡	93.40	93.30	96.20‡	-	-	84.40‡	-	-	85.90‡	-	-	-
UNIK (Yang et al. 2021a) arXiv'21	-	-	86.80	-	-	94.40	-	-	80.80	-	-	86.50	-	-	-
AdaSGN (Shi et al. 2021) arXiv'21	-	-	89.10	-	-	94.70	-	-	85.90	-	-	86.80	-	-	-
Yang's (Yang et al. 2021b) ECCV'21	88.00	-	-	94.90	-	-	-	-	-	-	-	-	-	-	-
MS-G3D (paper) (Liu et al. 2020a) CVPR'20	89.40	90.10	91.50	95.00	95.30	96.20	-	-	86.90	-	-	88.40	35.80	35.44	38.00
MS-G3D (code)* (Liu et al. 2020a) CVPR'20	88.77	89.59	90.67	94.88	94.86	95.82	82.35	84.86	86.42	84.14	86.79	87.98	35.74	34.77	37.23
STF (ours)	**91.34**	**91.09**	**92.47**	**96.46**	**96.51**	**96.86**	**85.06**	**86.80**	**88.85**	**86.40**	**88.86**	**89.92**	**38.20**	**37.56**	**39.87**

Table 2: Performance comparison of skeleton-based action recognition in top-1 accuracy (%). *: For MS-G3D (Liu et al. 2020a), the publicized code (Liu et al.) we use as our baseline has lower accuracy than what was reported in their paper. Annotations: J: joint; B: bone; "-": results not provided in the reference. The methods requiring more inputs than $J+B$: "⋆": (Song et al. 2020a) uses 3 streams; "‡": (Cheng et al. 2020; Li et al. 2020) use $J+B+J$ motion+B motion; "◇": (Shi et al. 2020a) uses spatial-temporal, spatial, slow-temporal, and fast-temporal streams; "†": (Song et al. 2020b) uses $J+B$+velocity.

Experiment: Scarce Data

method \ p	10	20	25	100
MS-G3D (code)	72.01	79.11	81.91	88.77
STF	**72.73** (↑0.72)	**80.77** (↑1.66)	**84.27** (↑2.36)	**91.34** (↑2.57)

Table 4: The accuracy (%) using joint input modality on the NTU-60 dataset under the x-sub setting. We use p% of the randomly sampled training data from the NTU-60 dataset.

Experiment: Dataset Shifting

- Train on NTU-60, tune last FC layer and test on NTU 120

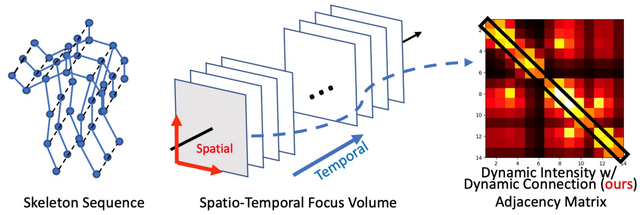
method	MS-G3D	STF
accuracy	82.34	**83.38**

Spatio-Temporal Focus

Extract Spatio-Temporal Focus in GradCam manner

$$Q = ReLU\left(\sum_c \left(\frac{1}{Z}\left(\sum_t \sum_v \frac{\partial \hat{y}}{\partial X^l_{ctv}}\right) X^l_{ctv}\right)\right)$$

where $\hat{y}$ is the category prediction score at class y, c, t, v denote the channel, temporal and spatial dimension of the input feature map X^l_{ctv} respectively. Z is the normalization factor of the spatio-temporal dimension.

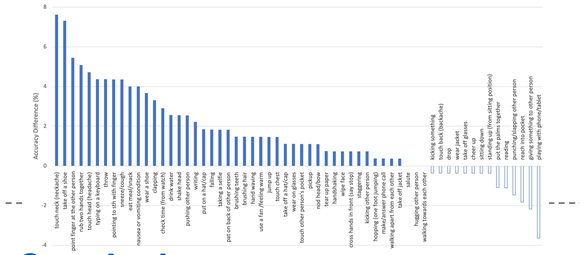


Experiment: Ablation Study

method	MS-G3D	L_c	L_d	L_{G_k}	L_e	accuracy	Δ
M_1	✓					88.77	—
M_2	✓	✓				89.33	↑0.56
M_3	✓	✓	✓			89.92	↑1.15
M_4	✓	✓	✓	✓		90.65	↑1.88
M_5: STF	✓	✓	✓	✓	✓	**91.34**	↑2.57

Table 5: Ablation study of top-1 accuracy (%) using joint only modality on the NTU-60 dataset under the x-sub setting. Δ shows the accuracy improvement over the baseline, MS-G3D (Liu et al. 2020a).

Experiment: Classwise improvement



Conclusion

- We design the novel STF module that generates dynamic connection **topology** and **intensity**
- We propose three loss terms defined on the gradient-based spatio-temporal focus to guide the classifier
  - *when and where to look at*
  - *distinguish confusing classes*
  - *optimize the stacked STF modules.*
- SOTA performance on three benchmarks, scarce data and dataset shifting settings.

University at Buffalo The State University of New York

MITSUBISHI ELECTRIC Changes for the Better | MITSUBISHI ELECTRIC RESEARCH LABORATORIES, INC