Global Co-occurrence Feature Learning and Active Coordinate System Conversion for Skeleton-based Action Recognition

Sheng Li, Tingting Jiang, Tiejun Huang, Yonghong Tian

NELVT, Department of Computer Science, Peking University, China

INTRODUCTION

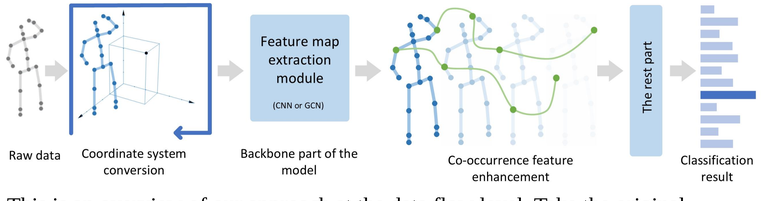
Motivation & Background:

1) Most skeleton-based action recognition methods do not fully explore the spatio-temporal co-occurrence, and many methods use separate strategies for spatial and temporal feature extraction.
2) On the other hand, the current method lacks some common modules compatible with multiple methods.

Idea & Contribution:

1) We propose a new spatio-temporal co-occurrence feature enhancement method. It can extract feature across spatial-temporal domain.
2) We design an active coordinate system transformation that can better align the skeleton data for action recognition. And it is compatible with most CNN-based or GCN-based methods.

Pipeline:

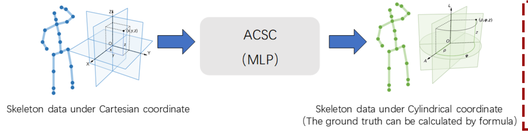


This is an overview of our approach at the data flow level. Take the original skeleton data as input, then pass ACSC and STUFE, and finally complete the classification.

METHOD

Active Coordinate System Conversion (ACSC):

Step1 (pretrain):

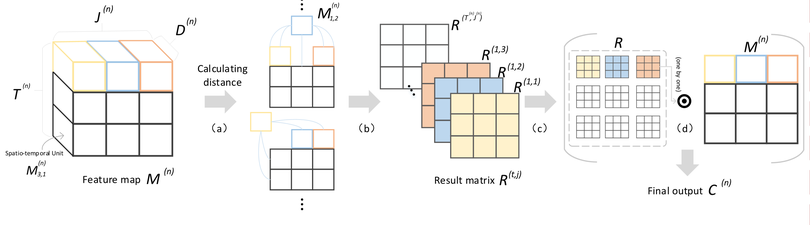


Pre-trained ACSC so that it can fit the transformation process from Cartesian coordinates to Cylindrical coordinates.

Step2 (End-to-end training):

Use the ground truth of the action label to supervise ACSC to complete the final training.

Spatio-temporal-unit Feature Enhancement (STUFE):



First step a) is to calculate the distance between every two spatio-temp units. Then all the results can be obtained after the step c) . Finally, the global co-occurrence feature enhancement is completed by self-attention mechanism according to the distance.

EXPERIMENT

NTU-RGB+D:

Method	CS	CV
Lie Group [25] *CVPR 2014*	50.1	52.8
H-RNN [4] *CVPR 2015*	59.1	64.0
Deep RNN [22] *CVPR 2016*	59.3	64.1
Deep LSTM [22] *CVPR 2016*	60.7	67.3
PA LSTM [22] *CVPR 2016*	62.9	70.3
ST-LSTM [18] *ECCV 2016*	62.9	70.3
STA-LSTM [24] *AAAI 2017*	73.4	81.2
Visualization CNN [20] *PR 2017*	76.0	82.6
VA-LSTM [33] *ICCV 2017*	79.4	87.6
Temporal Conv [12] *CVPRW 2017*	74.3	83.1
Clips + CNN + MTLN [10] *CVPR 2017*	79.6	84.8
Skepxels [17] *arXiv 2017*	81.3	89.2
HCN [15] *IJCAI 2018*	86.5	91.1
RHCN [Described in Sec. 4.2]	84.2	90.7
3D-POSE-S2 [21] *CVPR 2018*	82.4	86.7
ST-GCN [30] *AAAI 2018*	81.5	88.3
SR-TSL [30] *ECCV 2018*	84.8	92.4
motif-GCNs [28] *AAAI 2019*	84.2	90.2
STGR-GCN [14] *AAAI 2019*	**86.9**	92.3
RHCN + ACSC + STUFE	**86.9**	**92.5**

SBU Kinect Interaction :

Method	Accuracy (%)
Raw Position [32] *CVPRW 2012*	49.7
Joint feature [8] *ICMEW 2014*	86.9
CHARM [16] *ICCV2015*	86.9
H-RNN [4] *CVPR 2015*	80.4
ST-LSTM [18] *ECCV 2016*	88.6
Co-occurrence-LSTM [37] *AAAI 2016*	90.4
STA-LSTM [24] *AAAI 2017*	91.5
ST-LSTM + Trust Gate [18] *ECCV 2016*	93.3
VA-LSTM [33] *ICCV 2017*	97.6
RHCN + ACSC + STUFE	**98.7**

Top-1 accuracies on NTU-RGB+D and SBU Kinect Interaction benchmarks. We both archived high performance.

Training time comparison:

Method	training time (s)	increments(%)
RHCN	7023	baseline
RHCN+ACSC	7213	+2.71%
RHCN+STUFE	7721	+9.94%
RHCN+ACSC+STUFE	7962	+13.37%

Little increase in model's training time when ACSC and STUFE are added.

Ablation study:

Methods	Accuracy (%)			
	X=ST-GCN		X=RHCN	
	SBU	NTU	SBU	NTU
X	94.3	81.5	97.4	84.2
X + CCS	94.4	81.5	97.5	84.3
X + ACSC	94.8	81.7	97.7	84.9
X + STUFE	95.2	82.3	98.3	86.1
X + ACSC + STUFE	95.6	82.5	98.7	86.9

Ablation study on the SBU Kinect Interaction dataset and NTU-RGB+D dataset CS benchmark. CCS refers to the cylindrical coordinate system.

Related Video Link: