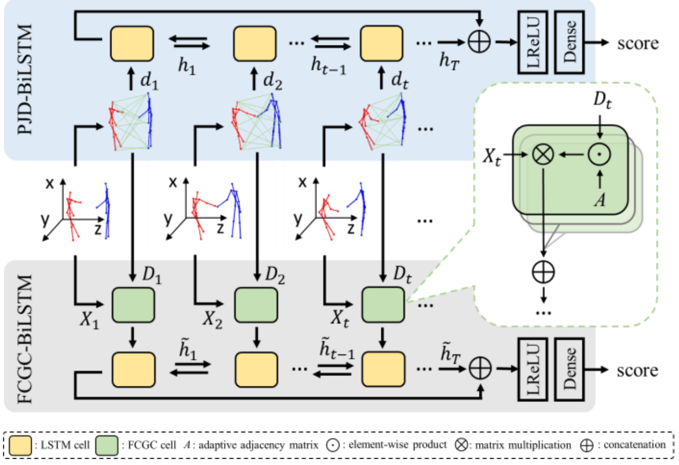
A Two-Stream Recurrent Network for Skeleton-based Human Interaction Recognition

Qianhui Men*, Edmond S. L. Ho†, Hubert P. H. Shum‡, Howard Leung*
*Department of Computer Science, City University of Hong Kong
†Department of Computer and Information Sciences, Northumbria University
‡Department of Computer Science, Durham University

ICPR 2020
25th INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION
Milan, Italy 10 | 15 January 2021

INTRODUCTION

- Goal: To recognize human-human interactions based on skeleton data from 2D or 3D joint locations.
- Weakness of Existing Work:
  - Lack of effective spatial modeling among joints.
  - Heavily rely on features within Individual characters.
- Motivation:
  - Exploring valuable *mutual information* between characters.
  - Investigating inner correlations among joints using *graph*.
  - Learning the spatial proximity with *pairwise geometric features* in the graph representation.



SCORE FUSION

- Late Fusion: To take advantage of different discriminative abilities of the two streams by combining their prediction scores.
- Objective: To highlight the lower entropy in the probability distributions of the two network classifiers, since it indicates higher confidence of the predicted class; and to hold back the less discriminative predictions (larger entropy) in the meanwhile.
- Method: The final prediction score is weighted from both streams. Specifically, α_n gives the degree of confidence towards the n-th network stream (here $n = 1$ or 2):

$$\alpha_n = 1 - \frac{\sum_{k=1}^{K} P_n(y_k|\mathbf{X},\Theta_n)\log\left(P_n(y_k|\mathbf{X},\Theta_n)\right)}{\sum_{m=1}^{N}\sum_{k=1}^{K} P_m(y_k|\mathbf{X},\Theta_m)\log\left(P_m(y_k|\mathbf{X},\Theta_m)\right)},$$

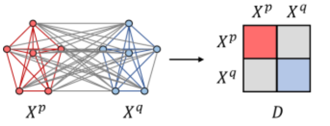
where $P_n(y_k|\mathbf{X},\Theta_n)$ is the k-th classification score of the interaction sample $\mathbf{X}$ under network parameter set Θ.

CONTRIBUTION

- A pairwise joint distance-based network (PJD-BiLSTM) that models the *explicit interaction patterns* from discriminative geometric features.
- A fully-connected graph convolution network (FCGC-BiLSTM) that quantifies the spatial proximity of interaction from joint positions to extract the *implicit correlations* among joints.
- A *late fusion* algorithm that takes advantage of both networks.
- State-of-the-art recognition performance on *3D interaction dataset* and comparable on RGB videos with *2D key joints*.

NETWORK STRUCTURE

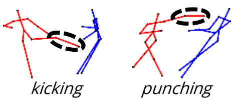
- Fully-Connected Mesh: Connecting any of the two joints of the interaction to capture features within two characters (X^p and X^q); Converting the pairwise distances into a weight matrix D.



Here is a simplified humanoid skeletal structure with five joints.

- PJD-BiLSTM: To learn the explicit spatial-temporal dependency by modeling the joint pair correlations. PJD of a joint pair X_i^p and X_j^q at frame t is calculated by:

$$D_t\left(X_i^p, X_j^q\right) = \frac{1}{\exp\left(\left\|X_i^p - X_j^q\right\|_t\right)}.$$



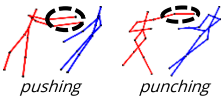
By modeling joint pair-level correlations, this stream is better at discriminating interactions with distinct patterns, such as *kicking* and *punching*.

- FCGC-BiLSTM: To capture the implicit correlations by representing the interaction as graph, where the skeletal joints of the two characters are graph nodes.

Given $\mathbf{X}_t = [X^p; X^q]|_t$ as input with C channels, the graph convolution operation inside FCGC cell under coefficients W is formed as:

$$W \star_g \mathbf{X}_t = \sigma\left(\left(\bigoplus_{c=1}^{C} (A \odot D_t)\mathbf{X}_t^c\right)W\right).$$

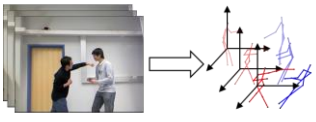
  - The joint connectivity A is adaptive to increase the flexibility of the graph representation.
  - PJD feature D_t is incorporated as auxiliary information to support the spatial proximity in the graph structure.



By modeling joint-level correlations, this stream is able to tell interactions with subtle differences, such as *pushing* and *punching*.

EXPERIMENT

Evaluation on 3D Interactions (SBU Interaction Dataset)



Method	Acc. (%)
Joint Features [1]	80.3
Clips+CNN+MTLN [2]	93.5
LSTM+FA+VF [3]	95.0
PJD-BiLSTM	94.0
FCGC-BiLSTM	95.1
PJD+FCGC	**96.8**

Evaluation on Key Joints of 2D RGB Videos (UT-Interaction Dataset)

Modality	Method	Acc. (%)
RGB	HR [4]	88.4
	PKM [5]	93.3
RGB+skeleton	PA-DRL [6]	**96.7**
skeleton	PJD-BiLSTM	91.9
	FCGC-BiLSTM	92.7
	PJD+FCGC	**94.4**

Comparisons of Confusion Matrix of Two Streams

approach	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
depart	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
push	0.00	0.00	0.00	0.93	0.00	0.02	0.00	0.05
shake hands	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.17
hug	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
exchange	0.00	0.00	0.00	0.00	0.00	0.00	0.97	0.03
punch	0.05	0.00	0.00	0.16	0.02	0.00	0.02	0.74

(a) PJD-BiLSTM on 3D

approach	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
depart	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	0.95	0.00	0.00	0.00	0.00	0.05
push	0.00	0.02	0.00	0.95	0.00	0.00	0.00	0.02
shake hands	0.00	0.00	0.00	0.00	0.83	0.00	0.00	0.17
hug	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
exchange	0.00	0.00	0.00	0.00	0.05	0.00	0.93	0.02
punch	0.00	0.00	0.00	0.08	0.00	0.00	0.02	0.90

(b) FCGC-BiLSTM on 3D

shake hands	1.00	0.00	0.00	0.00	0.00	0.00
hug	0.00	1.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	0.95	0.00	0.05	0.00
point	0.00	0.00	0.00	1.00	0.00	0.00
punch	0.00	0.00	0.10	0.00	0.75	0.15
push	0.00	0.10	0.00	0.00	0.10	0.80

(a) PJD-BiLSTM on 2D

shake hands	1.00	0.00	0.00	0.00	0.00	0.00
hug	0.00	1.00	0.00	0.00	0.00	0.00
kick	0.00	0.00	0.90	0.00	0.00	0.10
point	0.00	0.00	0.00	1.00	0.00	0.00
punch	0.00	0.00	0.05	0.00	0.80	0.15
push	0.00	0.00	0.00	0.00	0.15	0.85

(b) FCGC-BiLSTM on 2D

REFERENCE

[1] Two-person interaction detection using body-pose features and multiple instance learning, CVPRW'12
[2] A new representation of skeleton sequences for 3d action recognition, CVPR'17
[3] Attention-based multiview re-observation fusion network for skeletal action recognition, TMM'18
[4] A hierarchical representation for future action prediction, ECCV'14
[5] Poselet key-framing: A model for human activity recognition, CVPR'13
[6] Part-activated deep reinforcement learning for action prediction, ECCV'18