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

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.

CONTRIBUTION

- **Fully-Connected Mesh**: Connecting any of the two joints of the interaction to capture features within two characters (x<sup>i</sup> and x<sup>j</sup>). Converting the pairwise distances into a weight matrix D.
- **PID-BiLSTM**: To learn the *explicit spatial-temporal dependency* by modeling the joint pair correlations. PID of a joint pair x<sup>i</sup> and x<sup>j</sup> at frame t is calculated by:

\[
D_t(x^i_t, x^j_t) = \frac{1}{\exp(||x^i_t - x^j_t||)}
\]

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

**NETWORK STRUCTURE**

Given X = [x<sup>i</sup>, x<sup>j</sup>], as input with C channels, the graph convolution operation inside FCGC cell under coefficients W is formed as:

\[
X^* = W_t X = \sigma((A + D)_{t} X) W_t
\]

- The joint connectivity A is adaptive to *increase the flexibility* of the graph representation.
- PID feature D<sub>t</sub> is incorporated as auxiliary information to *support the spatial proximity* in the graph structure.

- **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<sub>n</sub> gives the degree of confidence towards the n-th network stream (here n = 1 or 2):

\[
\alpha_n = 1 - \frac{\sum_{k=1}^{N} \sum_{m=1}^{M} \log p(y^i_k | x, \theta_n) \log p(y_j^j_k | x, \theta_n)}{\sum_{k=1}^{N} \sum_{m=1}^{M} \log p(y^i_k | x, \theta_n) \log p(y_j^j_k | x, \theta_n)}
\]

where p(y<sup>i</sup><sub>k</sub> | x, \theta<sub>n</sub>) is the k-th classification score of the interaction sample x under network parameter set \theta.

**EXPERIMENT**

- **Evaluation on 3D Interactions (SBU Interaction Dataset)**
- **Evaluation on Key Joints of 2D RGB Videos (UT-Interaction Dataset)**
- **Comparisons of Confusion Matrix of Two Streams**

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc. (%)</th>
<th>Method</th>
<th>Acc. (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Joint Features</td>
<td>80.3</td>
<td>Modality</td>
<td></td>
</tr>
<tr>
<td>Clips+CNN+MTLN</td>
<td>93.5</td>
<td>RGB</td>
<td>88.4</td>
</tr>
<tr>
<td>LSTM+FA+VF</td>
<td>95.0</td>
<td>PKM</td>
<td>93.3</td>
</tr>
<tr>
<td>PJD-BiLSTM</td>
<td>94.0</td>
<td>RGB+skeleton</td>
<td>96.7</td>
</tr>
<tr>
<td>PJD+FGGC</td>
<td>96.8</td>
<td>PJD-BiLSTM</td>
<td>91.9</td>
</tr>
</tbody>
</table>

REFERENCES

[1] Two-person interaction detection using body-pose features and multiple instance learning, CVPRW13
[2] A hierarchical representation for future action prediction, ECCV14
[5] Attention-based multiview re-observation fusion network for skeletal action recognition, TMM18
[6] Part-activated deep reinforcement learning for action prediction, ECCV18

