Two-Person Interaction Detection Using Body-Pose Features and Multiple Instance Learning

Kiwon Yun
Jean Honorio, Debaleena Chattopadhyay, Tamara L. Berg, Dimitris Samaras

Stony Brook University
Indiana University
Human Activity Recognition

- Single-person activity
  eg. walking, running, waving, boxing, etc

[ C. Schuldt et al, ICPR’04 ]
Human Activity Recognition

- **Simple**
  - Single-person activity
  - eg. walking, running, waving, boxing, etc
  - I. Laptev et al., ICCV’03

- **Complex**
  - Daily living activity
  - eg. answering phone, eating, drinking, etc
  - [ R. Messing et al., ICCV’09 ]
  - Interactions
  - eg. kicking, punching, hand-shaking, etc
  - [ M. S. Ryoo + J. K. Aggarwal, ICCV’09 ]
Why interaction?

- Complex and non-periodic
- The study of causal relationships between two people or body parts could help extend the understanding of human motion.
- Necessary for a number of applications (eg. surveillance systems)
Outline

• Background
  – Motivation
  – Problem statement
  – Related work

• Dataset

• Real-time interaction detection
  – Body-pose features
  – Result

• Whole action sequence classification
  – Multiple instance learning
  – Result

• Conclusion and future work
• Recently, the rapid development of depth sensors provides a rather powerful human motion capture technique with low cost in real-time.
• This enables us to once again explore the feasibility of skeleton based features for activity recognition.
Problem statement

• Recognize human-human interactions performed by two persons from a single depth camera.
• Skeleton based body-pose features
• Two problems
  • Real-time activity detection
    – Choose activity category frame-by-frame in real-time.
  • Activity classification
    – Choose activity category given a segmented sequence.
Related work

- Interaction datasets
  - Video
  
  ![Video Image]

  [M. S. Ryoo et al., ICCV’09]  [A. Patron-Perez et al., BMVC’10]  [H. Kuehne et al., ICCV’11]

  - Motion capture
    - CMU Graphics Lab Motion Capture Database
    - Human Motion Database (HMD) [G. Guerra-Filho et al. FG’11]
Related work

- Kinect activity datasets
  - Single-person activity / gestures
    - [MSR Action 3D, CVPR'10 workshops]
    - [ChaLearn Gesture Challenge’11]
  - Daily activity
    - [B.Ni et al. ICCV’11 workshops]
    - [J. Sung et al., AAAI’11 workshops]

interaction dataset using Kinects?
Dataset

- We collected eight different human-human interactions performed by two persons using a Microsoft Kinect sensor.
We collected eight different human-human interactions performed by two persons using a Microsoft Kinect sensor. These include:

- approaching, departing, pushing, kicking, punching, exchanging an object, hugging, and shaking hands.
Dataset

- 7 participants, 21 sets (two-actor set), ~300 interactions
- It is composed of manually segmented videos for each interaction
Challenges

• Complex non-periodic actions
• Action-Reaction pattern
• Intra-class variation & Inter-class variation
• Skeletons contain incorrect tracking and noise.
Body-pose features

• Skeleton based features.


• Three types of features
  – Joint features (joint distance & joint motion)
  – Plane features (plane & normal plane)
  – Velocity features (velocity & normal velocity)

Body-pose features

Joint distance

Joint motion

Plane

Normal plane

Velocity

Normal velocity
Real-time interaction detection

How many frames do we need to detect an action?

Schindler, K. and Van Gool, L. CVPR’08

• Very short sub-sequences (1-7 frames), are sufficient for basic action recognition
• Can we recognize complex actions with features extracted within a very small time window?
Real-time interaction detection

- All body-pose features are computed within $W=3$ (20ms)
- Each feature vector is considered as a single instance in training
- SVM classifiers, 5-fold cross-validation
Real-time interaction detection

<table>
<thead>
<tr>
<th>Features</th>
<th>Average accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Raw position</td>
<td>0.497 ± 0.0480</td>
</tr>
<tr>
<td>Joint distance</td>
<td>0.793 ± 0.0276</td>
</tr>
<tr>
<td>Joint motion</td>
<td>0.802 ± 0.0390</td>
</tr>
<tr>
<td>Plane</td>
<td>0.612 ± 0.0282</td>
</tr>
<tr>
<td>Normal plane</td>
<td>0.723 ± 0.0333</td>
</tr>
<tr>
<td>Velocity</td>
<td>0.442 ± 0.0393</td>
</tr>
<tr>
<td>Normal Velocity</td>
<td>0.349 ± 0.0193</td>
</tr>
<tr>
<td>Joint features (Figure 2a &amp; 2b)</td>
<td>0.803 ± 0.0399</td>
</tr>
<tr>
<td>Plane features (Figure 2c &amp; 2d)</td>
<td>0.738 ± 0.0192</td>
</tr>
<tr>
<td>Velocity features (Figure 2e &amp; 2f)</td>
<td>0.484 ± 0.0387</td>
</tr>
<tr>
<td>Joint features + Plane features</td>
<td>0.790 ± 0.0349</td>
</tr>
<tr>
<td>Joint features + Velocity features</td>
<td>0.802 ± 0.0357</td>
</tr>
<tr>
<td>Velocity features + Plane features</td>
<td>0.744 ± 0.0201</td>
</tr>
<tr>
<td>All features</td>
<td>0.790 ± 0.0331</td>
</tr>
</tbody>
</table>

Figure: Detection performance (± standard deviation)
Window size = 3 (0.2 sec, real-time), SVM classifier, 5-folds cross validation
Real-time interaction detection

Confusion matrix of Joint features

<table>
<thead>
<tr>
<th></th>
<th>Approaching</th>
<th>Departing</th>
<th>Kicking</th>
<th>Pushing</th>
<th>ShakingHands</th>
<th>Hugging</th>
<th>Exchanging</th>
<th>Punching</th>
</tr>
</thead>
<tbody>
<tr>
<td>Approaching</td>
<td>0.92</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.02</td>
<td>0.01</td>
</tr>
<tr>
<td>Departing</td>
<td>0.00</td>
<td>0.97</td>
<td>0.01</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Kicking</td>
<td>0.04</td>
<td>0.01</td>
<td>0.85</td>
<td>0.03</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.05</td>
</tr>
<tr>
<td>Pushing</td>
<td>0.05</td>
<td>0.00</td>
<td>0.03</td>
<td>0.73</td>
<td>0.02</td>
<td>0.06</td>
<td>0.02</td>
<td>0.09</td>
</tr>
<tr>
<td>ShakingHands</td>
<td>0.02</td>
<td>0.00</td>
<td>0.02</td>
<td>0.05</td>
<td>0.75</td>
<td>0.01</td>
<td>0.12</td>
<td>0.04</td>
</tr>
<tr>
<td>Hugging</td>
<td>0.04</td>
<td>0.04</td>
<td>0.00</td>
<td>0.22</td>
<td>0.00</td>
<td>0.61</td>
<td>0.02</td>
<td>0.06</td>
</tr>
<tr>
<td>Exchanging</td>
<td>0.01</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>0.07</td>
<td>0.00</td>
<td>0.85</td>
<td>0.04</td>
</tr>
<tr>
<td>Punching</td>
<td>0.07</td>
<td>0.00</td>
<td>0.02</td>
<td>0.16</td>
<td>0.02</td>
<td>0.01</td>
<td>0.02</td>
<td>0.70</td>
</tr>
</tbody>
</table>
Real-time interaction detection

Confusion matrix of Joint features

[GT: Punching]

Pushing

Approaching | Departing | Kicking | Pushing | ShakingHands | Hugging | Exchanging | Punching
---|---|---|---|---|---|---|---
Approaching | 0.92 | 0.00 | 0.02 | 0.02 | 0.00 | 0.02 | 0.01
Departing | 0.00 | 0.97 | 0.01 | 0.00 | 0.00 | 0.01 | 0.00
Kicking | 0.04 | 0.01 | 0.85 | 0.03 | 0.00 | 0.00 | 0.01
Pushing | 0.05 | 0.00 | 0.03 | 0.73 | 0.02 | 0.06 | 0.02
ShakingHands | 0.02 | 0.00 | 0.02 | 0.05 | 0.75 | 0.01 | 0.12
Hugging | 0.04 | 0.04 | 0.00 | 0.22 | 0.00 | 0.61 | 0.02
Exchanging | 0.01 | 0.00 | 0.01 | 0.01 | 0.07 | 0.00 | 0.85
Punching | 0.07 | 0.00 | 0.02 | 0.16 | 0.02 | 0.01 | 0.70
Real-time interaction detection

Confusion matrix of Joint features
Whole sequence classification

... hugging? ... hugging?
We use Multiple Instance Learning (MIL) in a boosting framework to handle irrelevant frames in the training data.

We follow the MILBoost formulation in [P. Viola et al. NIPS’06]

The algorithm assigns a higher positive weight on a subset of instances, and these instances dominate subsequent learning.
Multiple Instance Learning
Whole action sequence classification

Set 1: original sequence
Set 2: earlier 5 frames + original sequence + later 5 frames
Whole action sequence classification

• We have shown that the MILBoost classifier outperforms SVMs if there exist uninformative frames in training data.

<table>
<thead>
<tr>
<th>Classifier</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Performance decrease</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear SVMs</td>
<td>0.876</td>
<td>0.687</td>
<td>-0.189</td>
</tr>
<tr>
<td>MILBoost</td>
<td>0.911</td>
<td>0.873</td>
<td>-0.038</td>
</tr>
</tbody>
</table>

Set 1: the original recorded sequence
Set 2: earlier 5 frames + the original recorded sequence + later 5 frames
Summary

• We have created a new dataset for two-person interaction using a Microsoft Kinect sensor.
• Geometric relational features based on distance between all pairs of joints within very small window outperformed other features for real-time interaction detection on noisy 3D skeleton data.
• Moreover, we have shown that the MILBoost classifier outperforms SVMs if there exist uninformative frames or sub-actions in training data.
Discussion and future work

• Extend the dataset?
  – More activity categories.
  – Different viewpoints / Multiple Kinects
• Better representations for interaction recognition?
• Better models? (HMMs, Graphical models, etc.)
• Data fusion (video + depth + skeleton)
• Can we investigate causal relationship between body parts of two actors, where one actor moves, and the other reacts?
Questions?