3D Human Action Recognition by Shape Analysis of Motion Trajectories on Riemannian Manifold

Maxime Devanne¹, Hazem Wannous¹, Stefano Berretti², Pietro Pala², Mohamed Daoudi¹, Alberto Del Bimbo²

¹ University Lille 1 / Télécom Lille
² MICC / University of Florence

REIMS IMAGE - CORESA’14, 26-27 Novembre 2014
Context

- Release of new depth cameras
  - Microsoft Kinect, Asus Xtion Pro Live
  - Depth image in addition to the color image
  - Skeleton estimated from depth image*

- Promising applications
  - Pose estimation
  - Hand gesture recognition
  - Human action recognition

* Shotton et al. CVPR 2011
Motivations

- **Goal:**
  - Recognition of the action performed by the subject in front of the sensor
  - Latency analysis

- **Challenges:**
  - Invariant to the position of the subject in the scene
  - Invariant to the speed of execution
State-of-the-art

- Skeleton-based approaches
  - Histograms of 3D joints [Xia et al. HAU3D 2012]
    - Spherical coordinates
    - Histograms of 3D joints

- EigenJoints [Yang et al. HAU3D 2012]
  - Pair-wise differences
State-of-the-art

- Depth Map-based approaches
  - 3D silhouettes [Li et al. HCBA 2010]
  - HOG on Depth Motion Maps [Yang et al. ACM Multimedia 2012]
  - Occupancy Pattern [Wang et al. ECCV 2012]

- Hybrid approach
  - Mining Actionlet [Wang et al. CVPR 2012]
    - Occupancy Pattern around each joint
Proposed Approach

Training

Spatio-temporal representation of sequences

Test

Shape analysis of trajectories

Shape space

K Nearest Neighbors

Action Class

Classification
Trajectories on action space

- Schema Trajectoires
Space-Time Representation

- Geometric invariance
  - Compute transformation between the first frame and a reference frame
    - Using Singular Value Decomposition and hip joints
  - Apply transformation to all frames of the sequence
Space-Time Representation

Action:
Sequence of skeletons with n joints

Feature vectors:
3D coordinates of each joints

Trajectory of the action in $\mathbb{R}^{3 \times n}$
Shape analysis of trajectories

- Square-root representation:

\[ q(t) = \frac{\beta(t)}{\sqrt{\|\beta(t)\|}} \]

[Joshi et al. CVPR 2007]

- Shape comparison between trajectories \( q_1 \) and \( q_2 \)
  - Compute geodesic distance between \( q_1 \) and \( q_2^* \):

\[ d_s(q_1, q_2^*) = \cos^{-1}(\langle q_1, q_2^* \rangle) \]
Shape analysis of trajectories

- Shape comparison between trajectories $q^1$, $q^2$
  - Find optimal $\gamma^*$ of $q^2$ wrt $q^1$ to obtain $q^2^*$
  - In practice dynamic programming is used
Classification

- Action recognition
  - K-Nearest Neighbor algorithm on shape space

- For a given trajectory \( q \):
  - Compute distance with all training trajectories
  - Keep the K nearest trajectories
  - Find the most frequent label
Classification

- Average Trajectories
  - Using Karcher Mean
  - Representative sequences
    - Per Action & Per Subject
Experiments

- Action Recognition

Recognition of the action

- Observational Latency Analysis

Recognition of the action
Experiments

- **MSR Action 3D dataset** [Li et al. HCBA 2010]
  - 20 actions performed by 10 subjects 2-3 times
    - Gaming action without any object

- **Used protocol**
  - Cross Subject Test:
    - 50 % of subjects for training, other 50% of subjects for test
Experiments

- MSR Action 3D dataset [Li et al. HCBA 2010]
- Comparison with the state-of-the-art’s works

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>EigenJoints</td>
<td>82.3 %</td>
</tr>
<tr>
<td>DMM</td>
<td>85.5 %</td>
</tr>
<tr>
<td>ROP</td>
<td>86.5 %</td>
</tr>
<tr>
<td>Actionlet</td>
<td>88.2 %</td>
</tr>
<tr>
<td>HON4D</td>
<td>88.9 %</td>
</tr>
<tr>
<td>DCSF</td>
<td>89.3 %</td>
</tr>
<tr>
<td>Our</td>
<td>88.3 %</td>
</tr>
<tr>
<td>Our + Karcher</td>
<td>92.1 %</td>
</tr>
</tbody>
</table>
Experiments

- MSR Action 3D dataset [Li et al. HCBA 2010]
  - Confounded actions
    - Hammer / draw tick, draw X

"Hammer" action
Experiments

- **UTKinect Dataset** [Xia et al. CVPRW 12]
  - 10 actions performed by 10 subjects 2 times (200 sequences)
    - Human-object interaction
    - Different points of view
    - Presence of occlusion

- Used Protocol
  - Leave-one-subject-out cross validation
Experiments

- **UTKinect Dataset** \[(Xia \ et \ al. \ CVPRW \ 12)\]
  - Comparison with the state-of-the-art

<table>
<thead>
<tr>
<th></th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hist. of 3D Joints [(Xia \ et \ al. \ CVPRW\ 12)]</td>
<td>90.9 %</td>
</tr>
<tr>
<td><strong>Our</strong></td>
<td>91.5 %</td>
</tr>
</tbody>
</table>
Experiments

- Observational Latency Analysis
  - Accuracy evolution with different portion of sequence
    - First n frames of the sequence

- UCF Kinect Dataset
  - 16 actions/16 subjects/5 times

- Comparison with state-of-the-art
  - Latency Aware Learning (LAL) [Ellis et al. IJCV 2013]
  - Two baseline solutions
    - Bag of Words (BoW)
    - Conditional Random Field (CRF)
Experiments

- Observational Latency Analysis

![Graph showing accuracy (%) vs maximum frames](image)
Conclusion

- A novel spatio-temporal representation of the action sequences
  - Trajectory in $\mathbb{R}^{60}$ based on the 3D position of joints
  - Shape comparison in a Riemannian manifold

- Competitive accuracies on public datasets
  - Failure cases
    - Repetition of an action
    - Human-object interaction
Ongoing work

- Investigate other descriptors based on depth to propose a hybrid approach
  - Especially in case of human-object interaction

- Analyse specific cases
  - The action is repeated more than once in the same sequence
  - Segmentation of the action
Thank you
Workshop Focus & Aims

Understanding human activities is a real problem, which needs an accurate acquisition of the movement sequence, consistent geometric representation of kinematics, dynamic modelling, and suitable learning process for motion identification. This workshop aims to bring together researchers from computer vision and machine learning communities, working together in a natural synergy and having an interest in using recent computing technologies to understand human, but also support him. In this way, we intend to present the recent vision-based algorithms in the related fields of static and temporal 3D data capture, modelling and representation, and their applications for social interactions. All aspects of 3D human sensing, such as detecting, tracking, motion and activity understanding will be addressed in this workshop. The covered topics include 3D pose estimation, human activity analysis, hand gesture analysis, body expression and body language. The workshop aims to provide an interactive platform for researchers to disseminate their most recent research results, discuss rigorously and systematically potential solutions and challenges, and promote new collaborations among researchers. In particular, the workshop is
References


Mean?
Experiments

- **Florence 3D Action Dataset** [Seidenari et al. CVPR 13]
  - 9 actions performed by 10 subjects 1-5 times
    - 215 sequences
    - Human-object interaction
    - Similarities between group of actions

- **Used Protocol**
  - Leave-one-subject-out cross validation
Experiments

- Florence 3D Action Dataset [Seidenari et al. CVPR 13]
- Comparison with the state-of-the-art

<table>
<thead>
<tr>
<th></th>
<th>Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>NBNN</td>
<td>82.0 %</td>
</tr>
<tr>
<td>Our</td>
<td>87.1 %</td>
</tr>
</tbody>
</table>