3D Human Video Retrieval: from Pose to Motion Matching

Rim Slama*, Hazem Wannous and Mohamed Daoudi
University of Lille / Telecom Lille1, France

* rim.slama@telecom-lille1.eu
Introduction

- Dynamic three-dimensional modeling of real-world objects using multiple cameras

- Long sequence of 3D videos: massive amounts of data
  - Browsing and searching for relevant information quickly become difficult
  - Need for 3D video segmentation system
  - Motion retrieval
  - Reuse of motion segments
Goal and Challenges

- **Goal**: Retrieve similar motion

- **Challenges**:
  1. Invariance to scale and local surface change
  2. Invariance to rotation and translation
  3. Invariance to sequence length
  4. Invariance to speed of the action execution
  5. Tolerance to changes in surface topology for similar poses
State of the art

Involvement of pose changes

- [Holte et al. ACM 2011]
  - ⊕ Thresholding step for segmentation

- [Pehlivan et al. CVIU 2010]
  - ⊕ Execution rate

Frame matching

- [Yamasaki et al. EURASIP 2007]
  - ⊕ topology
  - ⊕ Assumption of the motion nature

- [Huang et al. IJCV 2011]
  - ⊕ topology
  - ⊗ No sequence matching

- [Tung et al. PAMI 2012]
  - ⊕ Invariant to spatial location, scale
  - ⊗ topology, feature matching problem
Proposed Approach

EHC extraction

Modeling in shape space

Segmentation into clips

Trajectory

DTW

Clip matching

Motion retrieval
Pose matching

**Feature Extraction**

**Curve Extraction**

**Elastic shape analysis framework**

**Extremal Human Curves (EHC)** = geodesic paths between each two extremal points on the mesh.

**Feature points** = the points of a surface located at the extremity of its prominent components. [Tierny et al. IEEE 3DPVT 2009]
Pose matching

Shape analysis of curves

- We would like to take into account the elasticity of a human body
  - Bending
  - Stretching

- A parameterized curve on the body

\[ q : S^1 \rightarrow \mathbb{R}^3 \]
\[ q(t) \equiv \frac{\beta(t)}{\sqrt{\|\dot{\beta}(t)\|}} \]

\[ C = \{ q : I \rightarrow \mathbb{R}^3, \|q\| = 1 \} \subset L^2(I, \mathbb{R}^3) \]

[Srivastva et al. TPAMI 2011]
Pose matching

Shape curves similarity

Similarity measure: Length of geodesic path

Geodesic distance between \( q_1, q_2 \in \mathcal{C} \)

\[
d_c(q_1, q_2) = \cos^{-1}(\langle q_1, q_2 \rangle)
\]
Pose matching

3D human pose distance

- Best combination [Slama et al. FG13]
- Most stable ones and are sufficient to represent at best the body like a skeleton on the surface
Motion segmentation

- Splitting continuous sequence of motions into elementary segments

- Segments exhibit basic movements: meaningful ‘clips’

- Automatic segmentation of 3D video
  - without thresholding step
  - without assumption in the motion’s nature
**Segmentation approach**

- When human changes motion type or direction, the motion speed becomes small.

  Extrema detected on velocity curve should they be selected as segment points?

![Diagram of motion segmentation](image-url)
Result of clip segmentation: Walk

Motion segmentation

Distance between adjacent frames

Frames

0 to 250
Motion segmentation

Result of clip segmentation: Squat

Distance between adjacent frames vs Frames
Motion segmentation

Segmentation of slow and fast motion

Slow run

Fast run
Clip matching

- Clip matching: example based queries
- Clip: temporal sequence of human poses
- Human pose is an EHC representation

![Diagram of human poses]

- Curve extraction in video:
  1. Feature point tracking
  2. Curve extraction

15
Clip matching

Clip representation

• Representation of a sequence of extremal curves: a trajectory on the manifold of shape space.

• Comparing clips = comparing trajectories
  
  – ✗ Trajectories do not have the same length
  – Need for temporal alignment before computing distance

✓ Adapt DTW algorithm to features that reside on Riemannian manifold and use geodesic distance to compare similarity of curves.
**Clip matching**

**DTW on the manifold**

- **Clip 1**
- **Clip 2**

**Initial condition:** $\text{cost}(1, 1) = d(1, 1)$.

**DP-equation:**

$$\text{cost}(i, j) = \min \begin{cases} 
\text{cost}(i, j - 1) + d(i, j) \\
\text{cost}(i - 1, j - 1) + d(i, j) \\
\text{cost}(i - 1, j) + d(i, j) 
\end{cases}$$

**Warping window:** $j - r \leq i \leq j + r$.

**Time-normalized distance:**

$$D(\mathcal{A}, \mathcal{B}) = \frac{\text{cost}(n, m)}{\sqrt{n^2 + m^2}}$$
Experimental protocol

- 3D video dataset: i3DPost dataset
  - 13 motions: slow walk, fast walk, slow run, fast run, sprint, walk circle (left and right), run circle (left and right), walk in styles (cowboy, march, mickey) and complex motions (rock and roll, vogue dance)
  - 2 actors ➔ total of 26 actions

- Segmentation: 144 clips

- 14 classes of clips: walk Step (L/R)/ run Step (L/R)/walk march (L/R)/Rock dance-step1/dance-step2/(L/R) sprint step/ Mikey walk step (L/R)/cowboy walk step (L/R)

- Evaluation criteria:
  - Recall/Precision curve
  - 1st Tiers
  - 2nd Tiers
Example of some clips

walk

Run

Walk march

Rock and roll
Clip matching - Experiments

Retrieval results

Confusion matrix

<table>
<thead>
<tr>
<th></th>
<th>1st Tier(%)</th>
<th>2nd Tier(%)</th>
<th>E-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>DTW_EHC</td>
<td>79.26</td>
<td>93.44</td>
<td>53.27</td>
</tr>
</tbody>
</table>
Clip matching - Experiments

Example of clip retrieval

whole sequence

Clip 1

Clip 2

Similarity curve
**Conclusion**

**Summary**

- Extension of the use of EHC descriptor to motion retrieval
- Segmentation of the motion sequence by analysing of extrema of speed curve
- Representation of the motion clips by temporal trajectories on the manifold of open curve shape space
- Matching of the trajectories by DTW algorithm on the manifold

**Future works**

- Use HMM approach for action recognition
- Statistics on the manifold: mean pose, mean clip ...
- Hierarchical retrieval
Thank you
Q & A

rim.slama@telecom-lille1.eu