Cross-View Action Recognition from Temporal Self-Similarities

Imran Junejo, Emilie Dexter, Ivan Laptev, and Patrick Pérez

IRISA / INRIA Rennes, Campus universitaire de Beaulieu
35042 Rennes Cedex France

Objective

- Human actions recognition under view changes.

Related work

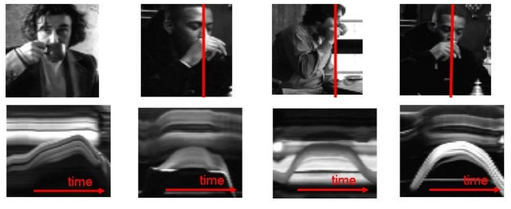
- Volumetric 3D reconstruction [Weinland et al. CVIU'06]
- Body part trajectories, projective geometry [Yilmaz and Shah ICCV'05], [Parameswaran and Chellapa IJCV'06]
- View-stable 2D trajectory features [Rao et al. IJCV'02].
- Projective geometry, no point correspondence [Wolf and Zomet IJCV'06].

Problems

- 2D/3D posture recovery is a difficult and generally unsolved problem.
- Direct extension of multiple view geometry methods to human actions is difficult due to the hard cross-view correspondence problem.
- Pure learning approach is difficult due to the limited number of action samples in different views.

Hypothesis

- View-invariance for non-rigid motion might be an easier problem compared to static scenes due to the additional time dimension.



This paper

- Cross-view action recognition under weak assumptions:
  - Only one test view
  - Different training view(s)
  - No 2D/3D reconstruction
  - No multi-view point correspondence
  - Assuming bounding box person localization

Self-Similarity Matrices (SSM)

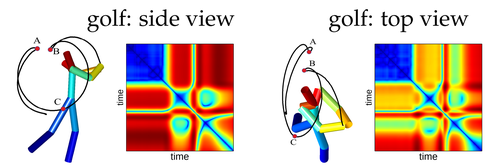
Temporal self-similarity matrix is defined as

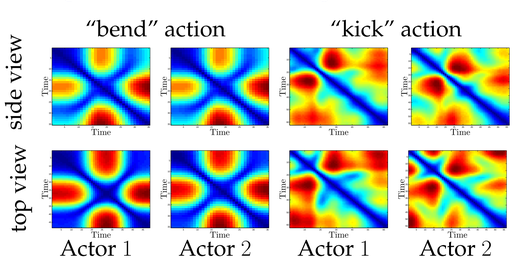
$$M = (m_{i,j})_{T\times T} \text{ with elements } m_{i,j} = k(x_i, x_j).$$

$m_{i,j}$ denotes the similarity of observations x_i, x_j at times i, j according to some kernel function k.

Trajectory-based SSM

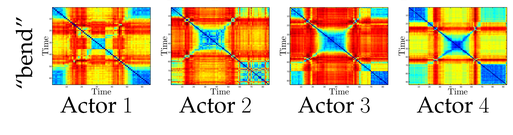
- $k(x_i, x_j) = \sum_p ||x_i^p - x_j^p||_2$.
- x_i^p: 2D point position on track p at time i.





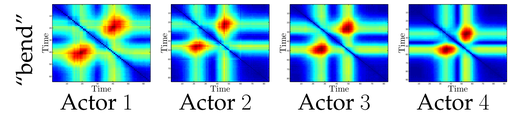
HoG-based SSM

- $k(x_i, x_j) = ||x_i - x_j||_2$.
- x_i: [Dalal&Triggs] person HoG descriptor.



Optical Flow based SSM

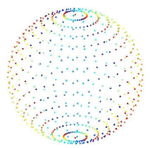
- $k(x_i, x_j) = ||x_i - x_j||_2$.
- x_i: OF components $\{v_x, v_y|v_x|v_y\}$ computed in person bounding box and concatenated into a vector.



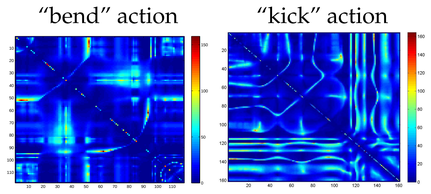
View Invariance Properties

- SSM "images" are stable under view changes

Experimental validation with dense view sampling



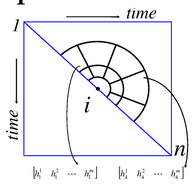
- Compute gradient orientation at each SSM point
- Estimate per-point orientation variance over views



average std: 17.4° average std: 20.93°

SSM-based Descriptor

- Local patch-based descriptor centered at each point on the diagonal
- Compute an 8-bin histogram of SSM gradients for each of the 11 blocks h_i^m



Action recognition: we represent each video as a bag of local SSM descriptors $H = (h_1, ..., h_n)$. Apply either Nearest Neighbor Classifier (BoF-NNC) or Support Vector Machine (BoF-SVM).

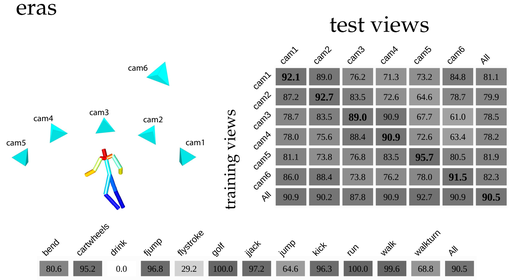
Temporal Multi-View Video Alignment

- Represent each SSM by sequences of local SSM descriptors H^1 and H^2.
- Align H^1 and H^2 with Dynamic Time Warping.

Action Recognition Results

CMU MoCap dataset (multi-view)

- Projected tracks for 13 joints, 12 action classes
- Simulated noise in joint tracking; six virtual cameras

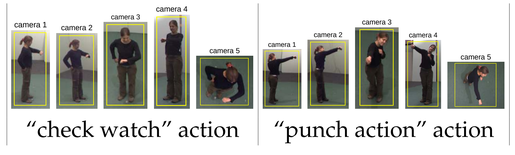


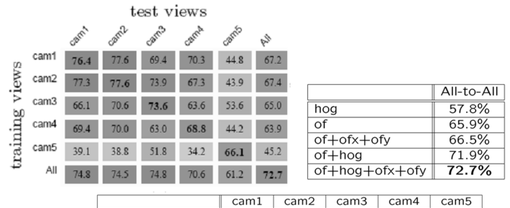
Weizman dataset (single view)

- 9 classes of actions, performed by 9 actors.
- NNC recognition accuracy 95.3% with SSM-pos and 94.6% with SSM-of-ofx-ofy-hog, compared to 92.6% [Ali et al. ICCV'07]

IXMAS dataset (multi-view)

- Dataset: 5 different views, 11 action classes, performed 3 times by 10 actors.





	All-to-All
hog	57.8%
of	65.9%
of+ofx+ofy	66.5%
of+hog	71.9%
of+hog+ofx+ofy	**72.7%**

	cam1	cam2	cam3	cam4	cam5
This paper	**76.4%**	**77.6%**	**73.6%**	**68.8%**	**66.1%**
Weinland et al. 3D	65.4%	70.0%	54.3%	66.0%	33.6%
Weinland et al. 2D	55.2%	63.5%	—	60.0%	—