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 ICCV’06]
- View-stable 2D trajectory features [Rao et al. IJCV’02]
- Projective geometry, no point correspondence [Wolf and Zomet ICCV’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

\[ M = (m_{ij})_{N \times N} \]

with elements

\[ m_{ij} = k(x_i, x_j) \]

\[ k(x_i, x_j) = \sum_{p=1}^{P} ||x_i^p - x_j^p||_2 \]

2D point position on track \( p \) at time \( t \).

Trajectory-based SSM
- \( k(x_i, x_j) = \sum_{p=1}^{P} ||x_i^p - x_j^p||_2 \)
- \( x_i^p \): 2D point position on track \( p \) at time \( t \).

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 \( \{v_x, v_y, v_z\} \) 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

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 \( h_i \)

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.

CMU MoCap dataset (multi-view)
- Projected tracks for 15 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-ofs-ofs-hog, compared to 92.6% [Ali et al. ICCV’07]

IXMAS dataset (multi-view)
- Dataset: 5 different views, 11 action classes, performed by 10 actors.