Learning Deformable Action Templates from Cluttered Videos

Benjamin Z. Yao and Song-Chun Zhu

Center for Image and Vision Science, Statistics Department, University of California Los Angeles.

1. Overview

Motivation
- Real-world actions exist in cluttered scenes (no easy segmentation and/or tracking).
- Previous methods either:
  1. Direct match between examples (too rigid) e.g. [Shechtman & Irani PAMI'07] [Yan Ke et al. ICCV07].
  2. Or discard spatial and/or temporal information (bag-of-words, too flexible) e.g. [Laptev et al. CVPR'08] [Jingen Liu et al. CVPR'09].
- We propose to learn a 3D (x, y, t) video template for each action class from a set of cluttered training videos (the CMU dataset).

Challenge: Spatial-temporal alignments?

Our approach at a glance
1. Initialize action template \( T_0 \) from an annotated example.
2. Find spatial-temporal alignment between \( T_0 \) and all other weakly labeled examples.
3. Adaptively select primitives to update \( T_0 \) by pooling over all aligned video frames.
4. Repeat 2, 3.

2. A generative model for action representation

1. Training videos and annotations:
   \[ V_m = (I_{m,1}^{(l)}, \ldots, I_{m,M}^{(l)}), m = 1 : M \]
   intensity image \( L \)-K optical flow
   \( \Lambda_m(t) \): image domain of target. It's a 3D (x, y, scale) function of time.
   \( w_m(\cdot) \): a time warping function that synchronizes each \( V_m \) with \( V_0 \).

2. Deformable shape template
   \[ u_m(w_m(t)) = \sum_{i=1}^{n} c_{m,i}B_{i,m,t} + u_m \]
   \( B_{i,m,t}, i = 1, \ldots, n \): Gabor wavelets
   \( c_{i,m,t}, i = 1, \ldots, n \): coefficients
   \( u_m \): unexplained residual image

3. Deformable motion template
   \[ f_m(w_m(t)) = \sum_{i=1}^{n} c_{i,m,t}O_{i,m,t} + u_m \]
   \( O_{i,m,t}, i = 1, \ldots, n \): local average filters of optical flow.

3. Semi-supervised learning with dynamic space-time warping

Using least square criterion:
\[
(c, B, c', O, A, w)^* = \arg\min_{c, B, c', O, A, w} \sum_{m=1}^{M} \int_{t=0}^{t_f} \left| \int_{t=0}^{t_f} \left| u_m(w_m(t)) - \sum_{i=1}^{n} c_{i,m,t}B_{i,m,t} \right|^2 dt + \lambda \| \Psi \Lambda_m(t) \| \right| \]

- \( \Theta \): template parameter (shared by all training examples)
- \( \Gamma \): alignments (different for each training example)
- Likelihood term
- Regularizer that enforces temporal smoothness of target’s trajectory

Semi-supervised learning procedure:
- \( \Gamma_0 \rightarrow \Theta_0 \rightarrow \Theta_1 \rightarrow \Gamma_1 \ldots \) where \( \Theta_0 \) is initialized from \( V_0 \)
- \( \Gamma_0 \rightarrow \Theta_1 \): we propose a dynamic space-time warping (DSTW) algorithm based on DTW.

4. Experiments

1. Learning action templates from the CMU action dataset [Y. Ke et al. ICCV'07]
2. Action detection on the CMU dataset
   - Time warping matrix. The white zig-zag line is the time warping path. Target in this video holds bending pose longer than normal.
3. Isolating shape and motion (CMU dataset)
   - AUC pick up 1 hand push b. jack 2 hand
     - shape: 0.53 0.46 0.68 0.34 0.44
     - motion: 0.21 0.29 0.12 0.31 0.27
     - both: 0.58 0.38 0.48 0.43 0.64
4. Action recognition AP on the KTH dataset
   - Ave. walk jog run box clap
     - 87.8% 95.3% 74.6% 81.3% 68.7% 93.8% 90.5%

Contributions
- A generative model of action with a likelihood defined directly on video frame intensities and optical flow fields.
- An efficient semi-supervised learning scheme that solves both adaptive primitive selection and spatial-temporal alignment while requiring only weak supervisions.

*Details please see [Y. Wu, Z. Si, H. Gong and S.C. Zhu, ICCV'07]
*Illustration of the DTW algorithm is adapted from [Eamonn Keogh, "Exact Indexing of Dynamic Time Warping", 2002]