Summarization and Indexing of Human Activity Sequences

Bi Song*, Namrata Vaswani**, Amit K. Roy-Chowdhury*

*EE Dept, UC-Riverside
**ECE Dept, Iowa State University
In order to summarize human activity sequences, we need to:

- Recognize the current activity from a set of known activity types
- Track using the activity’s model
- Detecting the change to the next activity
Applications

- Summarizing/annotating videos, e.g.
  - Sports videos, Training videos
  - Movies or documentaries
- Surveillance, e.g.
  - Recognizing activities in a parking lot
  - Shop surveillance
  - Airport surveillance
Example

Recognition

Tracking

Tracking Error
Outline

- Our Approach & Related Work

- State Space Model for Tracking
  - Shape dynamics for an activity
  - Transitioning to next activity model
  - Motion model
  - Observation model

- Change Detection (using ELL & TE), Recognition

- Experimental Results & Future Plans
‘Shape Activity’ (SA) approach

- Recognize activity using a few frames
  - Invariant scaled Euclidean camera motion

- Track with dynamic model of recognized SA
  - Separate dynamics of shape from that of camera motion
    (allows learning dynamics of activity with one camera and
    tracking with another, possibly moving, camera, where have a
    statistical model for camera motion dynamics)

- Keep detecting “change” from current SA model
  - Use a combination of “ELL statistic” & Tracking Error
  - ELL detects “gradual deviations” from current SA model
Closed-loop Framework for Simultaneous Tracking and Recognition

Gradual activity change: replace Tracking Error by “ELL statistic” [Vaswani, ACC’04]
Existing Work

- Condensation for gesture recognition
  - Only tracked affine deformations
  - Used a discrete state variable to model current gesture type: needed a set of particles for each gesture type

- DBN on discrete states, LGM on rest: track with RB-PF
  - Need to learn the model for discrete state dynamics
  - SLDS: Markov model for discrete state (special case of DBN)

- All above approaches: not invariant to camera motion
Statistical Shape Analysis [Dryden-Mardia’98]

- **Configuration**: set of K sampled contour locations
  - or B-spline control points or any other “feature points”
  - Represented as a K-dim complex vector, C

- **Shape**: C modulo translation, scale, in-plane rotation (scaled Euclid camera motion)
  - lies on a non-Euclidean space (hyper-sphere)
  - represent as tangent coordinate w.r.t. a “pole”

- **Motion**: trans, scale, in-plane rotation

Shape X Motion ←→ Configuration
‘Shape Activity’ Model

- Each activity represented by a “mean shape” and an AR model for deviations about the “mean”

- State = [motion, shape]
  - Motion = trans, scale, in-plane rotation
  - Dynamics: model for random camera motion
  - Shape = “tangent coordinates” of current shape w.r.t. the activity’s “mean shape” [Dryden-Mardia]
  - Dynamics: AR model on “tangent coordinates”
Modeling Human Activity Dynamics
--Details

Stationary and Piecewise-Stationary Shape Sequence on the shape manifold. In (a), we show a stationary sequence of shapes; at all times the shapes are close to the mean shape $\mu$ and hence the dynamics can be approximated in $T_\mu$ (tangent space at $\mu$). In (b), we show a piecewise-stationary sequence of shapes; the shapes move on the shape manifold.
State Space Model

- $\mu = \text{mean shape of current activity}$
- $s_t = \text{scale, } \theta_t = \text{rotation}$
- $v_t = \text{tangent coordinate of current shape w.r.t. } \mu$
- **Motion:** $\log s_t = \log s_{t-1} + n_{s,t}$, $\theta_t = \theta_{t-1} + n_{\theta,t}$
- **Shape:** $v_t = A v_{t-1} + n_{v,t}$

- Observed edge map either generated by predicted configuration, $C_t$ or by clutter [Condensation, IJCV’98]
  - Arrange $v_t$ as a complex vector
  - $z_t = (1-v_t^*v_t)^{1/2} \mu + v_t$, $C_t = z_t s_t e^{i\theta_t}$
At Activity Change Time…

- Track using a particle filter (PF)

- Get shape from tangent coordinate and current mean shape, $\mu$
  
  - $z_t = (1-v_t^*v_t)^{1/2} \mu + v_t$

- Compute its tangent coordinate w.r.t. $\mu_{\text{new}}$
  
  - $v_{t,\text{new}} = [I - \mu_{\text{new}} \mu_{\text{new}}^*] z_t e^{i \theta(z_t, \mu_{\text{new}})}$
Change Detection (slow): ELL [Vaswani, ACC’04]

- Tracking Error (TE) relies on “loss of track” of observations to detect changes

- Gradual changes get tracked by a particle filter (PF)

- ELL: measure of KLD b/w the posterior & the t step ahead prediction distribution of state (pdf of state given no observations)
  - uses “tracked part of change” to detect it, detects only gradual changes (which TE misses)
Computing ELL

\[ ELL_t^N = \frac{1}{N} \sum_{i=1}^{N} v_t^{(i)T} \Sigma_v^{-1} v_t^{(i)} + \text{constant} \]

- \( v_t^{i} \) = particles of tangent coordinate of current shape w.r.t. current activity’s mean

- \( ELL = \) posterior Expectation of the negative Log Likelihood of \( v_t \) being generated from \( N(0, \Sigma_v) \) which is the prior pdf of \( v_t \)
Change Detection (slow): ELL
ELL v/s TE for Slow Change

ELL detects faster than TE
Change Detection (Sudden): TE

Sudden activity changes will cause the PF with a large enough number of particles, and tuned to the dynamical model of a particular activity, to lose track. The tracking error (TE) will increase when the activity changes and this can be used to detect the change times. TE is calculated by

\[ TE = \sum_{k=1}^{K} \| q_k - f(q_k, G_t) \|^2 \]

- \( q_k \): \( k^{th} \) predicted landmark
- \( f(q_k, G_t) \): the nearest edge point of \( q^k \) along its norm direction
Change Detection (Sudden): TE
Recognition

This is done by projecting the observed shape in a frame onto the mean shape for each of the learned activities and choosing the one with the largest projection.

Specifically, given an observed image $I_t$, we label this frame as the activity that minimizes

$$\| \Gamma_t - s e^{j\theta} \mu_m + (a + jb) \|$$

Where $s$ -- scale, $\theta$ -- rotation, $a+jb$ -- translation
Experiments

- 10 human activities captured indoors
  - 1) bending across,
  - 2) walking toward camera and bending down,
  - 3) leaning forward and backward,
  - 4) leaning sideward,
  - 5) looking around,
  - 6) turning head,
  - 7) turning upper body,
  - 8) squatting,
  - 9) bending with hands outstretched,
  - 10) walking.
Summarizing and Index of Human Activity Sequences
Outdoor Sequence (ongoing work)

- **Walking with package in hand**
- **Walking toward camera**
Future Work

- Replace PF by PF-MT [Vaswani et al, ICASSP’06]
  - Local shape deformation per frame is small
  - PF-MT: IS only on motion, MT on shape

- Improving observation model by adding more features

- NSSA model for tracking, PSSA for recognition

- Tracking & activity analysis across a network of cameras

- Illumination invariant tracking

- Unsupervised Training: given a time seq. of landmarks, automatically segment it into pieces & learn dynamics