Action Detection in Complex Scenes with Spatial and Temporal Ambiguities

Yuxiao Hu§◊, Liangliang Cao§, Fengjun Lv†, Shuicheng Yan#, Yihong Gong†, and Thomas S. Huang§

§: University of Illinois at Urbana-Champaign
†: NEC Laboratories America
#: National University of Singapore
◊: Microsoft

Presented by Fengjun Lv
Objective and Motivation

- Detect human actions in complex real-world scenes

- Many applications:
  - Surveillance
  - Human Computer Interaction (HCI)
  - Video content retrieval
  - Retail solution
  - ...

- Action detection and tracking helps to discover customer shopping patterns
  - Which aisles did a shopper browse?
  - Which path did the shopper take?
  - Which products did the shopper touch?

- Our focus is on “touching of a product in a retail store”
  - Human detection and tracking are available
Challenges

- Complex environment
  - Cluttered background
  - Huge occlusions
  - Change of lighting

- Imperfect human detection and tracking

- Spatial and temporal ambiguities
  - Large variations in
    - Human appearance
    - Scale
    - Viewing angle
    - Execution style and speed
  - Hard to have a clear cut (spatial and temporal segmentation) of an action instance

- Non-repetitive, short duration
Related Work (1)

- Action recognition by global features
  - Motion history image [Davis and Bobick, 1997]
  - Oriented optical flow [Efros et al, 2003]
  - Action exemplars and volumetric features [Ke et al, 2007]
  - densely sampled local video patches [Boiman & Irani, 2005]
  - Mach filter response [Rodriguez et al, 2008]
  - Body configuration and evolvement [Zelnik-Manor & Irani, 2006], [Lkizler and Forsyth, 2007]

- Action recognition by spatial-temporal interest points
  - hierarchical structures [Jhuang et al, 2007]
  - implicit shapes [Wong et al, 2007]
  - local contexts [Wu et al, 2007]
  - 3D spin images [Liu et al, 2008]
  - 3D cube [Yuan et al, 2009]
Related Work (2)

- Action recognition by global features
  - image annotation [Yang et al, 2006]
  - face detection [Viola et al, 2005]
  - drug activity prediction [Zhang & Goldman, 2001]
Our approach: Simulated annealing + Multiple Instance Learning + Support Vector Machine (SMILE-SVM)

- Linear SVM to train an action classifier
- Multiple Instance Learning (MIL) to allow spatial and temporal ambiguities during training and testing
- Simulated Annealing to avoid local optimum by the traditional MIL.
The features

- On the cropped regions given by human detection and tracking across multiple frames
  - Motion feature: Motion History Image (MHI)
  - Appearance feature: Foreground Image (FI) after background subtraction
  - Local features: Histogram of Oriented Gradients (HOG)

- Feature vector is about 100-D
- Features are normalized
- Multiple features are concatenated
Multiple Instance

- Action instance $x_i$: a spatial-temporal segment (i.e. a 3D cube) of the whole video volume.
  - Positive $y_i = 1$: action did happen within the segment
  - Negative $y_i = -1$: otherwise

- Bag of instances $B_j = \{x_i\}$: a set of instance
  - Positive $Y_j = 1$: $\exists y_i = 1, \text{ for } x_i \in B_j$ or $\sum_{x_i \in B_j} \frac{y_i + 1}{2} \geq 1$
  - Negative $Y_j = -1$: $\forall y_i = -1, \text{ for } x_i \in B_j$
Multiple Instance Learning

- As in [Andrews et al, NIPS 02], SVM-based MIL is formulated as to minimize

$$\min_{y_i} \min_{w, b, \xi} \frac{1}{2} \|w\|^2 + c \sum_i \xi_i$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i, \xi_i \geq 0,$$

If $Y_j = 1$, \( \sum_{x_i \in B_j} \frac{y_i + 1}{2} \geq 1 \),

If $Y_j = -1$, \( \forall x_i \in B_j, y_i = -1 \)

- Difficult to minimize due to the large number of possible combinations of $y_i$. 

**MI-SVM**

- [Andrews et al, NIPS 02] uses a simple iterative algorithm called MI-SVM

**Algorithm 1 : MI-SVM**

1. Initialize a SVM model.
2. Do \( t = 1, 2, \ldots, n \)
3. Re-labete the instances in positive bags using SVM
4. Re-train SVM using the new labels.
5. Until converge.

- Easily to get stuck in local minimum
A new objective function

- maximize bag classification rate
- maximize margin of the classifier

\[ S = \max_{w,b,y_i} nc + \frac{k}{\|w\|^2} \]

nc is the bag classification rate

\[ 1/\|w\|^2 \] measures the SVM margin

Simulated Annealing is used to avoid local optimum
The algorithm

for each positive bag, label all its instances as positive: $y_i^0 = 1$
for each negative bag, label all its instances as negative: $y_i^0 = -1$
initialize $S_{opt} = 0$

for each annealing temperature $T$, do

for each iteration $t$, do

re-train SVM model $(w, b)$ for dataset $\{x_i, y_i^{t-1}\}$
re-label action instances and bags
compute bag classification rate $nc$ and obj. function $S_t$

if $S_t > S_{opt}$ record $S_{opt} = S_t$, $w_{opt} = w$, $b_{opt} = b$

if $P(S_t, S_{opt}, T) > \text{random}(0,1)$ /* $P = \exp \left( - \frac{S_{opt} - S_t}{T} \right)$ */
randomly choose $x_i$ with small confidence $|w^T x_i + b|$
flip the instance label of $x_i$: $y_i^t = - \text{sign}(w^T x_i + b)$
else
keep $y_i^t = y_i^{t-1}$
end if

decrease $T = \rho T$ /* $\rho = 0.8$ */

Output the SVM classifier with $(w_{opt}, b_{opt})$
## Results on CMU dataset

<table>
<thead>
<tr>
<th>Actions</th>
<th>jumping jacks</th>
<th>hand wave</th>
<th>two-handed wave</th>
<th>pick-up</th>
<th>pushing elevator button</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Prec. Recall</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Ke et al.</strong></td>
<td>0.75</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Ke et al.</strong></td>
<td>0.05</td>
<td><strong>0.63</strong></td>
<td><strong>0.85</strong></td>
<td><strong>1.0</strong></td>
<td>0.6</td>
</tr>
<tr>
<td><strong>Ke et al.</strong></td>
<td>0.05</td>
<td><strong>0.44</strong></td>
<td><strong>0.40</strong></td>
<td><strong>0.5</strong></td>
<td>0.3</td>
</tr>
<tr>
<td><strong>Ke et al.</strong></td>
<td>0.05</td>
<td><strong>0.2</strong></td>
<td><strong>0.2</strong></td>
<td><strong>0.1</strong></td>
<td><strong>0.5</strong></td>
</tr>
<tr>
<td><strong>Ke et al.</strong></td>
<td>0.05</td>
<td><strong>0.2</strong></td>
<td><strong>0.2</strong></td>
<td><strong>0.1</strong></td>
<td><strong>0.9</strong></td>
</tr>
</tbody>
</table>
Results on retail store dataset

- The dataset
  - Real video from a retail store in Tokyo
  - 1 hour long, 20 minutes for training
  - ~150 positive bags, ~50 for training
  - ~75k positive instances, ~25k for training
  - ~382 negative bags randomly selected from non-action tracking trajectories
  - ~113k negative instances, ~34k for training

- Speed
  - 10 temperatures
  - 50 iterations for each temperature
  - Training takes about 1 day
  - Testing is real-time
Results on retail store dataset
Results on retail store dataset
Results on retail store dataset
Results on retail store dataset
Summary & Future Work

- Multiple Instance Learning provides a natural way to handle spatial and temporal ambiguities in action detection problem
- Simulated Annealing helps!

- Add more sophisticated features
  - SIFT-like
  - Spatial-Temporal Interest Point

- Incorporate more sophisticated classifier
  - Non-linear SVM (\(\chi^2\), intersection, …)
  - Convolutional Neural Network

- Use faster learner
  - ASGD…
Thank you!

どうもありがとうございました!