Actom Sequence Models for Efficient Action Detection

LEAR – INRIA Grenoble

Adrien Gaidon
Zaid Harchaoui
Cordelia Schmid

Presentation by Benoit Massé
Introduction

- Video: Big Data
- Automatisation?
  - Semantic analysis
  - Retrieval

Problem:

Find if and when a specific action happen
State of the art

- Training
  - Define the action
  - Choose the features
  - Train

- Retrieval
  - Classification
  - Detection
State of the art

• Training
  – Define the action => Spatio-temporal extent
  – Choose the features => HoG, HoF, SP interest Point
  – Train => Bag-of-Feature

• Retrieval
  – Classification => SVM, Bayesian Network
  – Detection => ?
Actoms

- **Actom**: short atomic action
An actom has
- A location $t$
- A radius $r$

Actom descriptors: Set of visual words
- Bag of Features applied on HoG, HoF, Harris Interest points...
- Ponderated sum from $t - r$ to $t + r$
Interest of Actoms

• An **action** is composed of several **actoms**
  – New goal: find an ordered sequence of actoms
  – No temporal dependance inside an action
    • Gap between actoms
    • Overlap

• An action can be composed of very different parts
  => Classic methods compute the average
One Action = One Actom Sequence

- The radius $r_i$ of actom $i$ depends on its distance to the closer other actoms: $\min(t_i - t_{i-1}, t_{i+1} - t_i)$

- ASM : concatenation of actoms words

\[(X_{11}, \ldots, X_{1k}, X_{21}, \ldots, X_{2k}, X_{31}, \ldots, X_{3k})\]
Classification

• Given a new ASM \((x_{11}, \ldots, x_{nk})\), does it correspond to the trained action? (for instance: «drinking»)
  - Classic machine learning problem
  - Chosen solution: SVM
  - Including negative examples improves the classifier
Detection

- Given a video, find all the occurrences of the trained action. (for instance: « drinking »)

  For every 5 frames
  
  Set the current frame as the middle actom
  Generate candidates for other actoms
  Apply classification on the result
  
  End

  Delete non-maximal overlapping actions
Detection

Tricky step: Generating the other actoms

We must estimate the distance between actoms

- **Training**: Build the multivariate distribution \( \{t_{i+1} - t_i\} \)
  Remove the outliers

- **Estimation**: Try all the possible combinations
  (starting from the middle limit the error propagation)
Experiments

4 kind of actions
- Drinking
- Smoking
- Open a door
- Sit down

Criteria
- OV20 (20 % Overlap)
- OVAA (All Actions Overlap)

State of the art Comparison
- Bag of Features
- Bag of Features with a grid
- Other published methods
## Results

<table>
<thead>
<tr>
<th>Method</th>
<th>“Drinking”</th>
<th>“Smoking”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>matching criterion: OV20</td>
<td></td>
</tr>
<tr>
<td>DLSBP [3]</td>
<td>40</td>
<td>NA</td>
</tr>
<tr>
<td>LP [12]</td>
<td>49</td>
<td>NA</td>
</tr>
<tr>
<td>KMSZ [9]</td>
<td>54.1</td>
<td>24.5</td>
</tr>
<tr>
<td>BOF</td>
<td>36 (±1)</td>
<td>19 (±1)</td>
</tr>
<tr>
<td>BOF T3</td>
<td>44 (±2)</td>
<td>23 (±3)</td>
</tr>
<tr>
<td>ASM</td>
<td>57 (±3)</td>
<td>31 (±2)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>“Open Door”</th>
<th>“Sit Down”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>matching criterion: OV20</td>
<td></td>
</tr>
<tr>
<td>DLSBP [3]</td>
<td>13.9</td>
<td>14.4</td>
</tr>
<tr>
<td>BOF</td>
<td>12.2</td>
<td>14.2</td>
</tr>
<tr>
<td>BOF T3</td>
<td>11.5</td>
<td>17.7</td>
</tr>
<tr>
<td>ASM</td>
<td>16.4</td>
<td>19.8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>“Open Door”</th>
<th>“Sit Down”</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>matching criterion: OVAA</td>
<td></td>
</tr>
<tr>
<td>BOF</td>
<td>9.9</td>
<td>5.8</td>
</tr>
<tr>
<td>BOF T3</td>
<td>5.1</td>
<td>13.1</td>
</tr>
<tr>
<td>ASM</td>
<td>14.9</td>
<td>16.7</td>
</tr>
</tbody>
</table>
ASM gives better result than state-of-the-art, using the same data sets.

=> Actoms are particularly adapted for representing the temporal structure of actions into videos.