Localization using Faster R-CNN and Multi-Frame Fusion

Ryosuke Yamamoto, Nakamasa Inoue, Koichi Shinoda
Tokyo Institute of Technology
Outline

Motivation: detect an action concept “SittingDown”

Our method: Faster R-CNN + LSTM + Re-scoring

Annotation: Frame-wise annotation for SittingDown, Key-frame annotation for other concepts

Results:

2nd among 3 teams, best result at SittingDown
Motivation

• Localization task focuses not only on static objects, but also on action concepts
• We focus on SittingDown, one of action concepts
• How to distinguish between Sitting and SittingDown?
  → Dynamic information is important for precise detection
Our Method

- **Faster-RCNN** (Ren 2015)
  - Efficient object localization
- **LSTM** (Donahue 2015)
  - Precise action localization
  - Applied to SittingDown
- **Re-scoring** (Yamamoto 2015)
  - Multi-frame Score Fusion
  - Multi-Shot Score Boosting
Faster R-CNN (Ren 2015)

Efficient End-to-End object localization
1. Generate region proposals by a network
2. Predict scores for each region by using CNN features
Example CNNs:
   - ZF Net (Zeiler 2014) ← we use
   - VGG-16 (Simonyan 2014)
   - GoogLeNet (Szegedy 2015)
   - ResNet (He 2016)
Long Short-Term Memory (LSTM)

An LSTM layer is introduced to Faster R-CNN
- memorize long and short term information
- applied only to SittingDown
Multi-Frame and Multi-Shot (Yamamoto 2015)

- Multi-Frame Score Fusion
  Average pooling of scores over 5 frames in a shot
- Multi-Shot Score Boosting
  Add adjacent shot scores

\[
\text{score}^{\text{boost}}(r^t_i) = \text{score}(r^t_i) + \beta \max_j \frac{r^t_i \cap r^{t\pm1}_j}{r^t_i \cup r^{t\pm1}_j}
\]

\(r^t_i\): ith region in time \(t\); \(\beta\): multiplier
Key-Frame Annotations

Bounding-box annotation on the representative key-frame for each shot labeled as positive in collaborative annotation

<table>
<thead>
<tr>
<th>Concept</th>
<th># frames</th>
<th># boxes</th>
<th>Concept</th>
<th># frames</th>
<th># boxes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Animal</td>
<td>11,545</td>
<td>9,155</td>
<td>Inst.Musician</td>
<td>4,923</td>
<td>7,229</td>
</tr>
<tr>
<td>Bicycling</td>
<td>599</td>
<td>1,355</td>
<td>Running</td>
<td>945</td>
<td>1,394</td>
</tr>
<tr>
<td>Boy</td>
<td>1,848</td>
<td>2,492</td>
<td>SittingDown</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Dancing</td>
<td>2,118</td>
<td>5,199</td>
<td>Baby</td>
<td>898</td>
<td>895</td>
</tr>
<tr>
<td>ExplosionFire</td>
<td>2,483</td>
<td>2,402</td>
<td>Skier</td>
<td>320</td>
<td>521</td>
</tr>
</tbody>
</table>
I-Frame Annotations for SittingDown

- I-Frame annotation for SittingDown to train LSTM
- Annotation results
  
  * shots = 92
  * frames = 481
  * bounding-boxes = 515

* We found SittingDown in only 92 shots in the 3K shots labeled as positive in collaborative annotation
## Results

<table>
<thead>
<tr>
<th>ID</th>
<th>Method</th>
<th>RunID</th>
</tr>
</thead>
<tbody>
<tr>
<td>1*</td>
<td>Faster R-CNN + Multi-Frame Score Fusion</td>
<td>fusion</td>
</tr>
<tr>
<td>2*</td>
<td>1 + Multi-Shot Score Boosting</td>
<td>boost</td>
</tr>
<tr>
<td>3*</td>
<td>1 + LSTM(4096 units) for SittingDown</td>
<td>fusion.lstm</td>
</tr>
<tr>
<td>4*</td>
<td>2 + LSTM(4096 units) for SittingDown</td>
<td>boost.lstm</td>
</tr>
<tr>
<td>5</td>
<td>2 + LSTM(64 units) for SittingDown</td>
<td>(post exp.)</td>
</tr>
</tbody>
</table>

- 2nd among 3 teams
Results for SittingDown

Best result for SittingDown with run #2
LSTM with 4096 units (run #4) did not work
→ LSTM with 64 units (run #5) avoided over-fitting and worked in post submission experiment

<table>
<thead>
<tr>
<th>ID</th>
<th>Method</th>
<th>I-Frame F-score</th>
<th>Pixel F-score</th>
</tr>
</thead>
<tbody>
<tr>
<td>2*</td>
<td>Fusion + Boosting</td>
<td>0.63</td>
<td>0.22</td>
</tr>
<tr>
<td>4*</td>
<td>2 + LSTM (4096 units)</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>5</td>
<td>2 + LSTM (64 units)</td>
<td>11.96</td>
<td>4.51</td>
</tr>
</tbody>
</table>
SittingDown

Re-trained network with LSTM 64 units

Good cases

Bad cases

System output
Ground truth

Sitting down
Moving but not sitting down
Moving around a chair
Animal, Good Results

- Faster R-CNN
- Score Fusion
- Score Boosting

Cat (no movement)

Dog (walking)

System output
Ground truth
Animal, Bad Results

Faster R-CNN  Score Fusion  Score Boosting

Many animals

Bird (flying fast)

System output  Ground truth
Others

Faster R-CNN  Score Fusion  Score Boosting

Bicycling

Boy

System output  Ground truth
Others

Faster R-CNN  Score Fusion  Score Boosting

Dancing

ExplosionFire

System output  Ground truth
Others

Faster R-CNN  Score Fusion  Score Boosting

Instrumental Musician

Running

System output  Ground truth
Others

Faster R-CNN  | Score Fusion  | Score Boosting
---|---|---
Baby

Skier

System output  Ground truth
Conclusion & Future Work

- We proposed a localization system
  - Faster R-CNN + LSTM + Re-scoring
- Manual annotation
  - 31K bounding boxes
- Results
  - 2nd among 3 teams, best result at SittingDown
  - LSTM with 64 units was effective for SittingDown
- Future work
  - Find a better way to localize action