Speed/Accuracy Trade-offs for Object Detection from Video

Team Name: IC&USYD
Speaker: Jiankang Deng
Submission Brief

- Object detection from video with provided training data
  Rank 1# mAP: 81.8309%
  mAP: 80.8292% (2016 NUIST)

- Object detection from video with additional training data
  Rank 1# mAP: 81.9339%

- Object detection/tracking from video with provided training data
  Rank 1# mAP: 64.1474%
  mAP: 55.8557% (2016 CUVideo)

- Object detection/tracking from video with additional training data
  Rank 1# mAP: 64.2935%
VID Dataset

Class Number: 30
Training set: 4000 snippets
Validation set: 1314 snippets
Test set: 2000 snippets
VID Dataset Observation

1. Four leg mammal (18 classes)

2. Vehicle (7 classes)

3. Reptile (3 classes) and context related object, e.g. Bird(sky) and Whale(sea)
Challenges of VID

- Large appearance variation significantly affects prediction scores.
- Temporal information is important to improve the recall.
Submission 2015

- Object detection on each frame
- Object tracking from the high score frames
- Box regression and refinement
- False Positive suppression by context inference
Construct Correlation Filter \([1,2]\) on the Conv Maps

\[
r^* = \arg \min_r \sum_{i,j} \| r \cdot x_{i,j} - y(i, j) \|_2^2 + \lambda \| r \|_2^2, \quad y(i, j) = e^{-\frac{(i-W/2)^2+(j-H/2)^2}{2\sigma^2}}
\]

FFT

\[
R^k = \frac{Y \odot \overline{X}^k}{\sum_{k=1}^D X^k \odot \overline{X}^k + \lambda}
\]

Correlation Filter Update

\[
A^k_t = 0.3A_0 + (0.7 - \mu) A^k_{t-1} + \mu Y \odot \overline{X}^k
\]

\[
B^k_t = 0.3B_0 + (0.7 - \mu) B^k_{t-1} + \mu \sum_{k=1}^D X^k_t \odot \overline{X}^k
\]

\[
R^k_t = \frac{A^k_t}{B^k_t + \lambda},
\]

General tracking is object-oriented. VID task has class-specific prior.


Recent Works on VID

propagate deep features (ResNet101) by flow

Speed: 4.05 fps; mAP: 73.9%

Speed: 20.25fps; mAP: 73.1% [1]

Speed: 1.36fps; mAP: 76.3% [2]

Training Data and Crowded Status

1. Train model on DET data.
2. Predict the score of VID boxes.
3. Select positive examples [0.05, 0.9] from VID.
4. Remove redundant frames (low motion speed).
5. Balance training sample.

Loss: box classification; box regression; crowded status
Adaptive Frame Rate

Adaptive frame rate based on motion speed and appearance change

Birds of a feather flock together

Tree-based context model [1] (VOC 07)

Co-occurrence matrix on VID

Model Ensemble

Input Frames

ResNet101

ResNet269

Inception-ResNet v2

intact representation

Results

# Experimental Results

<table>
<thead>
<tr>
<th>Method</th>
<th>mAP (%) on the validation set</th>
<th>FPS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline: Single frame R-FCN (ResNet 101)</td>
<td>74.5</td>
<td>4.10</td>
</tr>
<tr>
<td>++ Adaptive Frame Rate Deep features propagation and aggregation by flow</td>
<td>76.8</td>
<td>15.4</td>
</tr>
<tr>
<td>++ Context inference (suppress FP)</td>
<td>77.6</td>
<td></td>
</tr>
<tr>
<td>++ Short tractlet combination and re-scoring (similar to seq-NMS)</td>
<td>80.7</td>
<td></td>
</tr>
<tr>
<td>++ Global stage-wise re-rank</td>
<td>82.4</td>
<td></td>
</tr>
<tr>
<td>Submission</td>
<td>mAP (%) on the test set</td>
<td></td>
</tr>
<tr>
<td>++ Ensemble ResNet 269 and Inception-ResNet v2</td>
<td>81.8309</td>
<td></td>
</tr>
</tbody>
</table>
Demo Video
Demo Video
Demo Video
Team member

Jiankang Deng$^1$ Yujiang Zhou$^1$ Baoshen Yu$^2$ Zhe Chen$^2$ Stefanos Zafeiriou$^1$ Dacheng Tao$^2$

1. Intelligent Behavior Understanding Group (ibug), Imperial College London, UK
2. UBTECH Sydney AI Centre, University of Sydney, Australia
Email: jiankangdeng@gmail.com