Tracking the Untrackable: Learning to Track Multiple Cues with Long-Term Dependencies

CS637 Paper Presentation
Alrik Firl
What is Tracking?

- Localize and associate object trajectories through a sequence of frames
  - Single / Multi-object
  - Multi / Moving camera
  - Online tracking
- Tracking-by-detection:
  - Associate new detections with tracked objects in other frames
- Important and old problem, many application areas
- Occlusions, noisy detections, appearance variation, matching ambiguity, etc

MPDNN on ETH-Linthescher (2D MOT 2015)
Tracking the Untrackable: High Level Overview

- Structure of RNNs to jointly reason over multiple cues
  - Appearance Model (A) -- $\phi_A$
  - Motion Model (M) -- $\phi_M$
  - Interaction Model (I) -- $\phi_I$
- Compute object feature vector $\phi_O$ via target LSTM (O) from $\phi_A$, $\phi_M$ and $\phi_I$
  - Similarity score for association
  - Hungarian algorithm for matching in MDP framework
- Online method

- What it is:
  - Method to compute track to object similarity
- What it isn’t:
  - Tracking method
- Motivation:
  - Model history of each cue
  - Combine multiple cues
Markov Decision Processes (MDP)

- Online MOT → decision making in MDP
  - Lifetime of track modeled with an MDP
  - Learning track similarity function → learning MDP policy
- Optical-flow based single-target tracking
  - Iterative Lucas-kanade w/ pyramids [2]
  - Forward-backward error for stability measure of track [3]
- Associate track to detections for lost state
  - Use LSTM features for similarity matrix
  - Associate with Hungarian
Appearance Model

- Model visual similarities between track and detection
  - Re-identification problem
  - Robust to occlusion, visual perturbations
- CNN (VGG16) computes feature extraction per frame
- LSTM computes object feature across frames
- Compute similarity of j to past appearance history of object i
  - Compute for all tracks i
- Pre-train w/ binary classifier (whether j corresponds to i)

\[ BB_i^1 \cdots BB_i^t \rightarrow \text{CNN} \rightarrow \text{H-D feature} \rightarrow \text{LSTM} \rightarrow \text{H-D feature} \phi_i \]

\[ BB_j^{t+1} \rightarrow \text{CNN} \rightarrow \text{H-D feature} \phi_j \]

\[ \{\phi_i, \phi_j\} \rightarrow \text{FC} \rightarrow \text{k-D feature} \]
Appearance Model CNN

- Pretrained VGG16 model
- Replace last FC layer w/ 500-D FC layer
- Fine-tune CNN on re-identification task (CUHK dataset) [7]
  - Siamese CNN
  - Train on 2DMOT2015 and CUHK03
    - + samples: same target in different frames
    - - samples: different targets across all frames
- Siamese CNN
  - Share weights
  - Good for computing similarity between different inputs

Siamese CNN re-identification results on CUHK03

<table>
<thead>
<tr>
<th>Method</th>
<th>Rank 1</th>
<th>Rank 5</th>
<th>Rank 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>FPNN [40]</td>
<td>19.9</td>
<td>49.3</td>
<td>64.7</td>
</tr>
<tr>
<td>BoW [91]</td>
<td>23</td>
<td>45</td>
<td>55.7</td>
</tr>
<tr>
<td>ConvNet [2]</td>
<td>45</td>
<td>75.3</td>
<td>95</td>
</tr>
<tr>
<td>LX [41]</td>
<td>46.3</td>
<td>78.9</td>
<td>88.6</td>
</tr>
<tr>
<td>MLAPG [42]</td>
<td>51.2</td>
<td>83.6</td>
<td>92.1</td>
</tr>
<tr>
<td>SS-SVM [88]</td>
<td>51.2</td>
<td>80.8</td>
<td>89.6</td>
</tr>
<tr>
<td>SI-CI [72]</td>
<td>52.2</td>
<td><strong>84.3</strong></td>
<td>92.3</td>
</tr>
<tr>
<td>DNS [87]</td>
<td>54.7</td>
<td>84.8</td>
<td>94.8</td>
</tr>
<tr>
<td>SLSTM [69]</td>
<td><strong>57.3</strong></td>
<td>80.1</td>
<td>88.3</td>
</tr>
<tr>
<td>Ours</td>
<td>55.9</td>
<td>81.7</td>
<td><strong>95.1</strong></td>
</tr>
</tbody>
</table>
Motion Model

- Independent motion of objects
  - Occlusions, dropped tracks (FN)
- Noisy detections
  - Nonlinear motion in image plane
- LSTM input: all velocities $v_i^{t_1...t_n} = (x_i^{t_1...t_n}, y_i^{t_1...t_n})$ of target $i$ → feature $\phi_i$
- $v_j^{t+1}$ → FC → H-D feature $\phi_j$
- $\{\phi_i, \phi_j\}$ → 2H-D feature $\phi$ → FC → $\phi_M$
- Output: k-D feature vector
- Trained as binary classifier matching $v_i^{t_1} ... v_i^{t_n}$ to $v_j^{t+1}$
Interaction Model

- Motion also influenced by surrounding objects
- Image: 15x15 uniform grid
  - Occupancy grid: 7x7 subwindow
  - Locations: bounding box centers
  - $O_{i}^{t_{1}...t_{n}}$: track i’s grid through time
  - $O_{i}^{t_{1}...t_{n}} \rightarrow$ LSTM $\rightarrow$ H-D feature $\phi_{i}$
  - $O_{j}^{t+1} \rightarrow$ FC $\rightarrow$ H-D feature $\phi_{j}$
  - $\{\phi_{i}, \phi_{j}\} \rightarrow$ FC $\rightarrow$ $\phi_{i}$ k-D feature
- trained as binary classifier matching $O_{j}^{t+1}$ to $O_{i}^{t_{1}...t_{n}}$

$$O_{i}^{t}(m, n) = \bigvee_{j \in \mathcal{N}_{i}} \mathbf{1}_{mn}[x_{j}^{t} - x_{i}^{t}, y_{j}^{t} - y_{i}^{t}]$$
Overall Architecture

- 1 LSTM per cue (A, M, I), gives features $\phi_A, \phi_M, \phi_I$
- 1 target LSTM (O) per track $t$
  - Computes similarity vector between track $t$ and detection $d$
  - Learns long-term dependencies across all cues in time
  - Input: 3k-D input $\{\phi_A, \phi_M, \phi_I\}$
  - Output: H-D $\rightarrow$ FC $\rightarrow$ similarity feature $\phi_O(t, d)$
  - Trained as binary classifier for matching track $t$ and detection $d$
Training Procedure and Dataset

1st phase: Pre-train LSTM models (A, M, I) as binary classification
- Softmax classifier, cross-entropy loss
  +: new detection matches prev. trajectory of target
  -: does not match
- Pre-train appearance CNN on re-identification task

2nd phase: Jointly train O, A, M, I
- Train O as binary classifier for associating detection to tracks

MOT15
- Pedestrian only
- Moving & stationary camera
- 11 training, 11 test

MOT16
- Multiclass (people, cars, motorcycles, bikes, etc)
- Moving & stationary camera
- 7 training, 7 test

Metrics:
- MOTA: tracking accuracy
- MOTP: tracking precision
- FAF: #false alarms / frame
- MT: ratio of mostly tracked objects
- ML: ratio of mostly lost objects
- FP: #false positives
- FN: #false negatives
- ID Sw.: #track ID switches
- Frag: #track fragmentations
- Hz: tracker speed (FPS)
Evaluation

- MDP Framework [1]
  - Use similarity metrics computed from network
  - 1 target (O) LSTM per target
  - 20% increase in MOTA over MDP similarity features
- MOT16
- Stanford Drone Dataset (SDD) [6]
<table>
<thead>
<tr>
<th>Tracker</th>
<th>Tracking Mode</th>
<th>MOTA↑</th>
<th>MOTP↑</th>
<th>MT↑</th>
<th>ML↓</th>
<th>FP↓</th>
<th>FN↓</th>
<th>IDS↓</th>
<th>Frag ↓</th>
<th>Hz ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SiameseCNN [36]</td>
<td>Offline</td>
<td>29.0</td>
<td>71.2</td>
<td>8.5%</td>
<td>48.4%</td>
<td>5,160</td>
<td>37,798</td>
<td>639</td>
<td>1,316</td>
<td>52.8</td>
</tr>
<tr>
<td>CNNTMC [71]</td>
<td>Offline</td>
<td>29.6</td>
<td>71.8</td>
<td>11.2%</td>
<td>44.0%</td>
<td>7,786</td>
<td>34,733</td>
<td>712</td>
<td>943</td>
<td>1.7</td>
</tr>
<tr>
<td>TSMLCDEnew [70]</td>
<td>Offline</td>
<td>34.3</td>
<td>71.7</td>
<td>14.0%</td>
<td>39.4%</td>
<td>7,869</td>
<td>31,908</td>
<td>618</td>
<td>959</td>
<td>6.5</td>
</tr>
<tr>
<td>JointMC [27]</td>
<td>Offline</td>
<td>35.6</td>
<td>71.9</td>
<td>23.2%</td>
<td>39.3%</td>
<td>10,580</td>
<td>28,508</td>
<td>457</td>
<td>969</td>
<td>0.6</td>
</tr>
<tr>
<td>TC. ODAL [5]</td>
<td>Online</td>
<td>15.1</td>
<td>70.5</td>
<td>3.20%</td>
<td>55.80%</td>
<td>12,970</td>
<td>38,538</td>
<td>637</td>
<td>1,716</td>
<td>1.7</td>
</tr>
<tr>
<td>RMOT [81]</td>
<td>Online</td>
<td>18.6</td>
<td>69.6</td>
<td>5.30%</td>
<td>53.30%</td>
<td>12,473</td>
<td>36,835</td>
<td>684</td>
<td>1,282</td>
<td>7.9</td>
</tr>
<tr>
<td>SCEA [79]</td>
<td>Online</td>
<td>29.1</td>
<td>71.1</td>
<td>8.9%</td>
<td>47.3%</td>
<td>6,060</td>
<td>36,912</td>
<td>604</td>
<td>1,182</td>
<td>6.8</td>
</tr>
<tr>
<td>MDP [75]</td>
<td>Online</td>
<td>30.3</td>
<td>71.3</td>
<td>13.00%</td>
<td>38.40%</td>
<td>9,717</td>
<td>32,422</td>
<td>680</td>
<td>1,500</td>
<td>1.1</td>
</tr>
<tr>
<td>TDAM [80]</td>
<td>Online</td>
<td>33.0</td>
<td>72.8</td>
<td>13.3%</td>
<td>39.1%</td>
<td>10,064</td>
<td>30,617</td>
<td>464</td>
<td>1,506</td>
<td>5.9</td>
</tr>
<tr>
<td>Ours</td>
<td>Online</td>
<td>37.6</td>
<td>71.7</td>
<td>15.8%</td>
<td>26.8%</td>
<td>7,933</td>
<td>29,397</td>
<td>1,026</td>
<td>2,024</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 5. Tracking performance on the test set of the 2D MOT 2015 Benchmark with public detections.

<table>
<thead>
<tr>
<th>Tracker</th>
<th>Tracking Mode</th>
<th>MOTA↑</th>
<th>MOTP↑</th>
<th>MT↑</th>
<th>ML↓</th>
<th>FP↓</th>
<th>FN↓</th>
<th>IDS↓</th>
<th>Frag ↓</th>
<th>Hz ↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>LINF1 [14]</td>
<td>Offline</td>
<td>41.8</td>
<td>74.8</td>
<td>11.60%</td>
<td>51.30%</td>
<td>7,896</td>
<td>99,224</td>
<td>430</td>
<td>963</td>
<td>1.1</td>
</tr>
<tr>
<td>MHT_DAM [31]</td>
<td>Offline</td>
<td>42.9</td>
<td>76.6</td>
<td>13.60%</td>
<td>46.90%</td>
<td>5,668</td>
<td>97,919</td>
<td>499</td>
<td>659</td>
<td>0.8</td>
</tr>
<tr>
<td>JMC [78]</td>
<td>Offline</td>
<td>46.3</td>
<td>75.7</td>
<td>15.50%</td>
<td>39.70%</td>
<td>6,373</td>
<td>90,914</td>
<td>657</td>
<td>1,114</td>
<td>0.8</td>
</tr>
<tr>
<td>NOMT [8]</td>
<td>Offline</td>
<td>46.4</td>
<td>76.6</td>
<td>18.30%</td>
<td>41.40%</td>
<td>9,753</td>
<td>87,565</td>
<td>359</td>
<td>504</td>
<td>2.6</td>
</tr>
<tr>
<td>OVBVT [6]</td>
<td>Online</td>
<td>38.4</td>
<td>75.4</td>
<td>7.50%</td>
<td>47.30%</td>
<td>11,517</td>
<td>99,463</td>
<td>1,321</td>
<td>2,140</td>
<td>0.3</td>
</tr>
<tr>
<td>EAMTF.pub [61]</td>
<td>Online</td>
<td>38.8</td>
<td>75.1</td>
<td>7.90%</td>
<td>49.10%</td>
<td>8,114</td>
<td>102,452</td>
<td>965</td>
<td>1,657</td>
<td>11.8</td>
</tr>
<tr>
<td>oCF [30]</td>
<td>Online</td>
<td>43.2</td>
<td>74.3</td>
<td>11.30%</td>
<td>48.50%</td>
<td>6,651</td>
<td>96,515</td>
<td>381</td>
<td>1,404</td>
<td>0.4</td>
</tr>
<tr>
<td>Ours</td>
<td>Online</td>
<td>47.2</td>
<td>75.8</td>
<td>14.0%</td>
<td>41.6%</td>
<td>2,681</td>
<td>92,856</td>
<td>774</td>
<td>1,675</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 6. Tracking performance on the test set of the MOT16 Benchmark with public detections.
References


