Tell Me What to Track
Qi Feng1, Vitaly Ablavsky1, Guorong Li2, Stan Sclaroff2, Qinxun Bai3
1Department of Computer Science, Boston University. 2University of Chinese Academy of Sciences. 3Hikvision Research America.

Contributions
Our major contributions are threefold:
1. A model that utilizes natural language specification to initialize tracking and to reset the tracker as needed;
2. An LSTM based language module that scores similarity between a given natural language description and visual features computed for regions in the image;
3. An LSTM network based tracker that performs tracking by detections.

Main Problem
We define our problem as a Single Object Tracking with Natural Language Specification.

Notations
The overall architecture is summarized in Figure 3. Here we pose the problem of single object tracking with NL query. Let \( I_1, \ldots, I_T \) denote the sequence of video frames, where \( T \) is the number of frames in a temporal window. Let \( F_{RPN}, F_{lang} \) be the function for RPN and Language network. Let \( X_t \) be proposed bounding boxes on frame \( t \). \( X_t \in B \), where \( B \) is the set of all possible bounding boxes.

Probabilistic View
For detection at frame \( t \), we wish to estimate \( \Pr[X_t|Q, I_t] \). Assume that \( I_t \) and NL \( Q \) are conditional independent given \( X_t \), following Bayes theorem, we have

\[
\Pr[X_t|Q, I_t] = \frac{\Pr[Q|I_t, X_t] \Pr[X_t]}{\Pr[Q|I_t]} (1)
\]

\[
\Pr[X_t|Q, I_t] = \frac{\Pr[Q|X_t] \Pr[X_t]}{\Pr[Q|I_t]} (2)
\]

\[
\Pr[X_t|Q, I_t] = \frac{\Pr[Q|I_t, X_t] \Pr[X_t]}{\Pr[Q|I_t]} (3)
\]

\[
\Pr[X_t|Q, I_t] = \frac{\Pr[Q|I_t] \Pr[X_t]}{\Pr[Q|I_t]} (4)
\]

\[
\Pr[X_t|Q, I_t] = \frac{\Pr[X_t|Q]}{\Pr[Q]} (5)
\]

Notice that \( \frac{\Pr[Q|I_t]}{\Pr[Q]} \) is a constant. Therefore, we have

\[
\Pr[X_t|Q, I_t] \propto F_{lang}(X_t, Q) \cdot F_{RPN}(X_t, I_t) (6)
\]

For each frame \( I_t \), the top \( N \) ranked proposals are returned as detections \( X_t^{lang} \).

Similarly, the tracker works as a sequential process which is to estimate

\[
\Pr[X_t|I_{1:t}, Q] \propto \Pr[X_t|I_{1:t-1}] \Pr[X_t|Q] (7)
\]

\[
\Pr[X_t|I_{1:t}, Q] \propto \Pr[X_t|I_{1:t-1}] \Pr[X_t|Q] (8)
\]

\[
\Pr[X_t|I_{1:t}, Q] \propto \Pr[X_t|I_{1:t-1}] \Pr[X_t|Q] (9)
\]

Results
Preliminary qualitative results from Figure 2 show that the proposed tracker is capable of handling occlusions and rapid motion changes by restarting the tracker with language detections. Quantitative results are summarized in Table 1.

<table>
<thead>
<tr>
<th>Model</th>
<th>AUC</th>
<th>MAttNet</th>
<th>Ours</th>
<th>Siams FC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>25.9</td>
<td>38.5</td>
<td>44.0</td>
<td>60.8</td>
</tr>
</tbody>
</table>

References