Robust Visual Tracking via Deep Discriminative Model

Heng Fan[1], Jinhai Xiang[2], Guoliang Li[2] and Fuchuan Ni[2]

[1]Department of Computer & Information Sciences, Temple University
[2]College of Informatics, Huazhong Agricultural University

1. Overview

Goal

- Develop a robust discriminative model for visual tracking.

Approach

- Design a deep discriminative model (DDM) by deep features with two sub-models, i.e., deep object-background model (DOBM) and deep object-distractor model (DODM).
- Generate object candidates with saliency proposal, which is able to help search target from a global area.
- Choose tracking result by evaluating each candidate with the DDMs.

2. Related Works

- Convolutional neural networks (CNNs) [1] have drawn extensive interests in computer vision [2, 3, 4] due to their powerfulness in feature extraction.
- Color features have been successfully utilized to develop discriminative model for tracking [5]. However, only using this (hand-crafted)color feature is less robust to deal with complex scenes in tracking.
- A simple method [6] used to estimate the saliency parts in image is able to generate object candidates from a global area, which helps to alleviate model drift problem.

3. The Proposed Method

Deep object-background model (DOBM):

Denote object region, background region and search region as *O*, *B* and *R*, respectively. Let F_{Ψ}^{R} be the histogram of deep feature extracted over $\Psi \in R$, $F_{\Psi}^{R}(b)$ the b^{th} bin of F, and $b_{\mathbf{x}}$ the b^{th} bin assigned to $R(\mathbf{x})$. Thus, we can obtain the object likelihood at location $\mathbf{x}$ by Bayesian

$$P(\mathbf{x} \in \mathcal{O}|O,B,b_{\mathbf{x}}) \approx \frac{P(b_{\mathbf{x}}|\mathbf{x} \in O)P(\mathbf{x} \in O)}{\sum_{\Psi \in \{O,B\}} P(b_{\mathbf{x}}|\mathbf{x} \in \Psi)P(\mathbf{x} \in \Psi)} \quad (1)$$

In particular, we can compute the likelihood terms by

$$P(b_{\mathbf{x}}|\mathbf{x} \in O) \approx \frac{F_O^R(b_{\mathbf{x}})}{|O|},\ P(b_{\mathbf{x}}|\mathbf{x} \in B) \approx \frac{F_B^R(b_{\mathbf{x}})}{|B|} \quad (2)$$

Where $|\cdot|$ is cardinality. Likewise, we can derive the prior probabilities $P(\mathbf{x} \in O)$ and $P(\mathbf{x} \in B)$ as follows:

$$P(\mathbf{x} \in O) \approx \frac{|O|}{|O|+|B|},\ P(\mathbf{x} \in B) \approx \frac{|B|}{|O|+|B|} \quad (3)$$

Substituting (2) and (3) into (1), we get the DOBM as follows

$$P(\mathbf{x} \in \mathcal{O}|O,B,b_{\mathbf{x}}) = \frac{F_O^R(b_{\mathbf{x}})}{F_O^R(b_{\mathbf{x}}) + F_B^R(b_{\mathbf{x}})}, \mathbf{x} \in (O \cup B) \quad (4)$$

For unseen pixels, i.e., $x \notin (O \cup B)$, their object likelihoods are set to 0.5.

Deep object-distractor model (DODM):

DODM is used to reduce the risk of model drift caused by similar distractors, and it is similar to DOBM except that the background region is replaced with a set of distracting regions D. Thus, similar to (4), the DODM is defined with

$$P(\mathbf{x} \in \mathcal{O}|O,D,b_{\mathbf{x}}) = \frac{F_O^R(b_{\mathbf{x}})}{F_O^R(b_{\mathbf{x}}) + F_D^R(b_{\mathbf{x}})}, \mathbf{x} \in (O \cup D) \quad (5)$$

Deep discriminative model (DDM):

Combining DOBM and DODM, we can get the DDM as follows

$$P(\mathbf{x} \in \mathcal{O}|b_{\mathbf{x}}) = \alpha P(\mathbf{x} \in \mathcal{O}|O,B,b_{\mathbf{x}}) + (1-\alpha)P(\mathbf{x} \in \mathcal{O}|O,D,b_{\mathbf{x}}) \quad (6)$$

where α is a pre-defined parameter. Using DDM, we can obtain an object-background confidence map, as shown in Fig. 1.

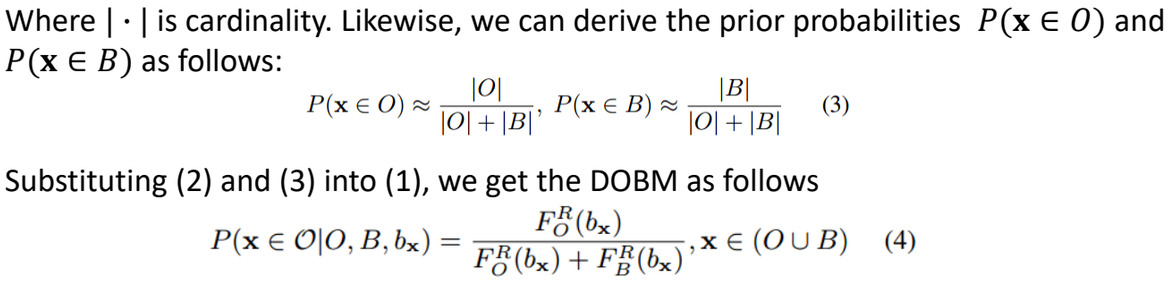

Figure 1: Illustration of the DDM.

Construct multiple DDMs:

Considering that both spatial information from lower layers and sematic information from higher layers benefit tracking, we construct multiple DDMs, and each of them corresponds one layer from VGG-16 network [4]. The final object-background confidence map P_{final} is

$$P_{final} = \sum_{l=1}^{L} w^l P^l(\mathbf{x} \in \mathcal{O}|b_{\mathbf{x}}) \quad (7)$$

where $P^l(\mathbf{x} \in \mathcal{O}|b_{\mathbf{x}})$ is the DDM of layer l, and w^l is its weight. Outside the search region, we set the confidence values of pixels to zeros.

Tracking:

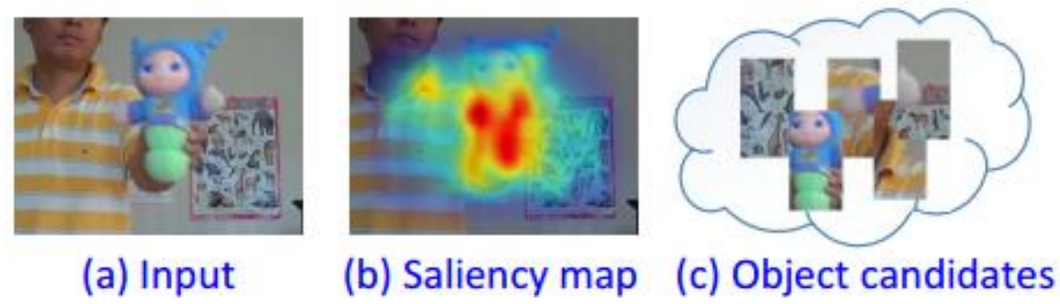

Figure 2: Illustration of generating candidates.

First, we generate a set of candidates denoted in frame t as $C^t = \{c_i^t\}_{i=1}^N$ (see Fig. 2), and we compute the confidence value for each candidate

$$V(c_k^t) = \sum_{(i,j) \in c_k^t} P_{final}^t(i,j) \quad (8)$$

The tracking result $\Phi^t = \underset{c_k^t}{\operatorname{argmax}}\, V(c_k^t)$. To adapt our tracker to changing appearance, we adopt a simple linear interpolation strategy to update mode.

4. Pipeline

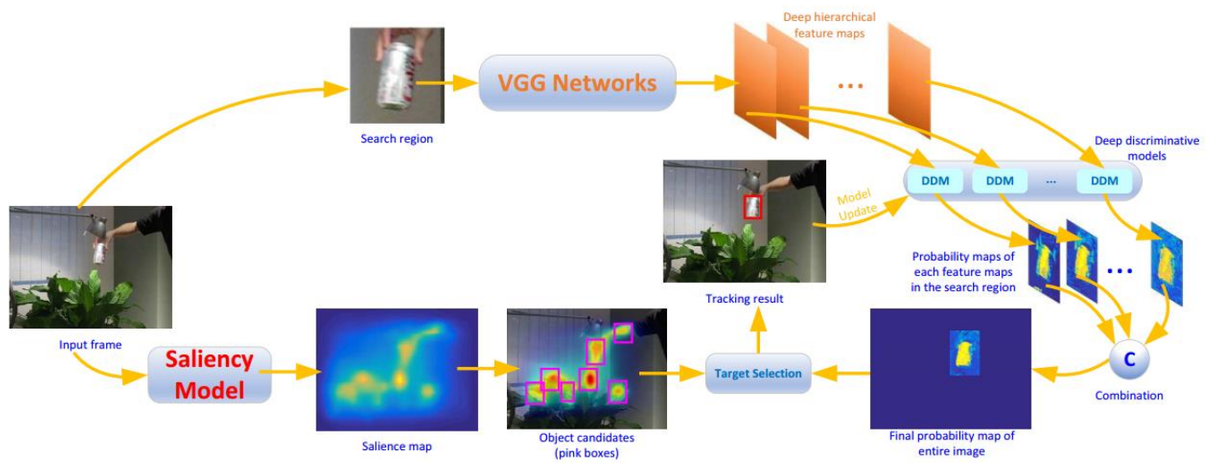

Figure 3: Tracking pipeline of the proposed method.

5. Results

We evaluate the proposed method on a large-scale tracking benchmark [7]. Fig. 4 and Fig. 5 show the results and comparisons with other methods.

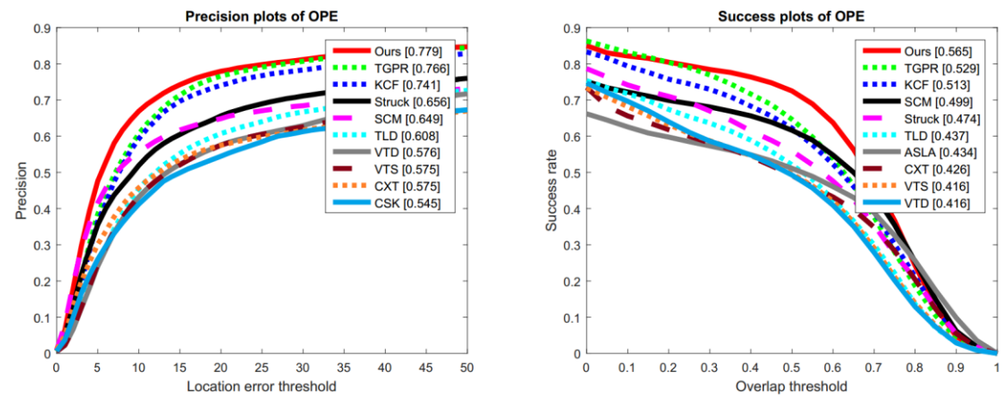

Figure 4: Comparisons of precision and success plots.

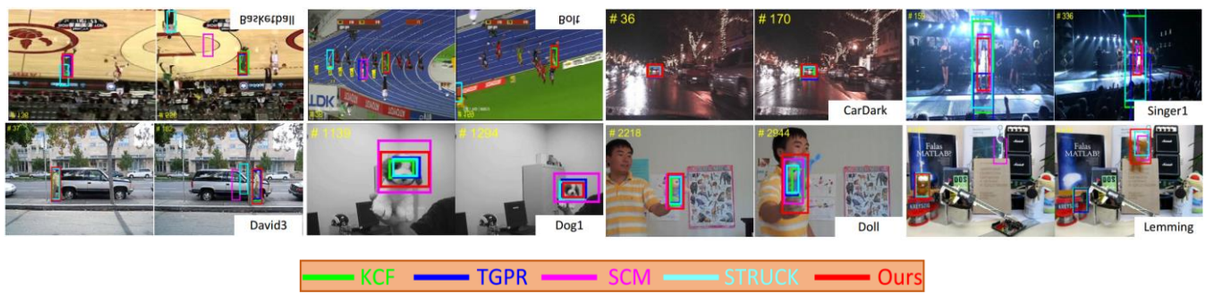

Figure 5: Qualitative results of five trackers on eight sequences.

6. Key References

[1] Y. LeCun, B. Boser, J. Denker, D. Henderson, R. Howard, W. Hubbard and L. Jackel, "Backpropagation applied to hand-written zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541-551, 1989.
[2] K. He, X. Zhang, S. Ren and J. Sun, "Delving deep into rectifiers: surpassing human-level performance on imagenet classification," in *CVPR*, 2015, pp. 1026-1034.
[3] A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012, pp. 1097-1105.
[4] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.
[5] H. Pessegger, T. Mauthner and H. Bischof, "In defense of color-based model-free tracking," in *CVPR*, 2015, pp. 2113-2120.
[6] J. Harel, C. Koch and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006, pp. 545-552.
[7] Y. Wu, J. Lim and M.-H. Yang, "Online object tracking: A benchmark," in *CVPR*, 2013, pp. 2411-2418.