RSINet: Rotation-Scale Invariant Network for Online Visual Tracking

Yang Fang (fangyang968@gmail.com), Geun-Sik Jo* and Chang-Hee Lee
Inha University, KAIST

Introduction

Siamese network trackers
- No model update and cannot learn target-specific variation adaptively
- Axis-aligned model contains extra noise
- Weak at rotation and scale estimation

Proposed RSINet tracker
- Model update adaptively and dynamically
- Object-aligned model without extra noise
- Tailored for rotation and scale estimation

Algorithm 1: Proposed RSINet Tracker.
Input: Pre-trained Network model M and Initial frame $I_0$ with annotation.
Output: Estimated target state $\bar{t}_t = (x_t, y_t, s_t, r_t)$; Updated model filters $h_t$.
while frame $t \leq \text{length(Video sequence)}$ do
    Feed new frame into Siamese network to predict new target state $(x_t, y_t, s_t, r_t)$.
    if $(t \geq 10)$ then
        Calculate spatio-temporal energy $\varepsilon$, in [9]
        if $\varepsilon \geq \kappa_{\varepsilon_0}$ then
            Derive steepest descend update rate $\alpha_\varepsilon$ [10]
            $\alpha \leftarrow \min\left(\frac{1}{\varepsilon}, \alpha_\varepsilon\right)$.
        end
    end
    Update tracking model filter $h_{t+1} = h_t + \alpha \nabla L(h_t)$.
end

Experimental evaluation

Conclusion

Proposed RSINet enables target-distractor model and rotation-scale model learning simultaneously. It keeps a good balance between tracking accuracy (0.604 on vot18) and running efficiency (45 FPS).