Spatial-sequential-spectral Context Awareness Tracking

西安交通大學 XI'AN JIAOTONG UNIVERSITY

Jianwu Fang, Zheng Li, and Jianru Xue

長安大學

1. Introduction

Visual context has formed a robust stimulation for visual perception. Spatio-temporal context in existing trackers sometimes shows weak reliability in visible light videos with poor quality. This paper proposes **3SContext** (select Spatial-Sequential-Spectral Context to track), a tracker which selects Spatial-Sequential-Spectral context to approximate the most discriminative power for target-background separation. The main property of this model is that the unreliability of spectral band in certain frames can be restrained, and make a truly reflection of the spectra discrimination. Besides, we handle the occlusion and scale estimation respectively by a trajectory regression and closed object contour. We provide the result that the proposed method can boost the performance significantly and outperforms many complex trackers on 50 videos while running at a real-time speed.

2. Problem Formulation

The tracking problem in this work is formulated as to estimate a rectangle r_t at time t in frame I_t, which gives the target location with a max-score obtained by our *3SContext* function:

$$r_t = \arg\max_{r_t \in I_t} f(\mathbf{M}(I_t, r); \mathcal{C}_t),$$

where $\mathcal{C}_t = \{\kappa_t^c, \varsigma_t^c, \rho_t^c\}$ is the context space containing the spatial context κ_t^c, sequential context ς_t^c, and spectral context ρ_t^c. $\mathbf{M}(\cdot)$ is a mapping function which bridges the image to target location. By that, $f(\mathbf{M}(I_t, r_t); \mathcal{C}_t)$ assigns a score to a rectangle window r_t in I_t in accordance with the context space.

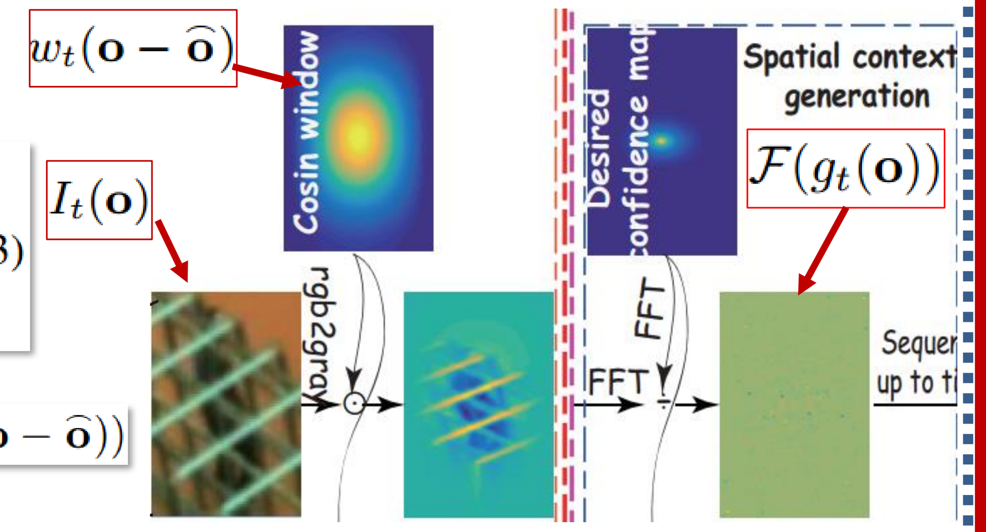
2.1. Inference

2.1.1. By κ_t^c

Spatial context models the relationship between the object location with its local surrounding region.

$$\begin{aligned} &f(\mathbf{M}(I_t, r); \kappa_t^c) \\ &= g_t(\mathbf{o}) \otimes (I_t(\mathbf{o}) w_t(\mathbf{o} - \bar{\mathbf{o}})) \\ &= \sum_{\mathbf{s} \in R_t} g_t(\mathbf{o} - \mathbf{s}) I_t(\mathbf{s}) w_t(\mathbf{s} - \bar{\mathbf{o}}) \end{aligned} \quad (3)$$

$$\mathcal{F}(f(\mathbf{M}(I_t, r); \kappa^c t)) = \boxed{\mathcal{F}(g_t(\mathbf{o}))} \odot \mathcal{F}(I_t(\mathbf{o}) w(\mathbf{o} - \bar{\mathbf{o}}))$$

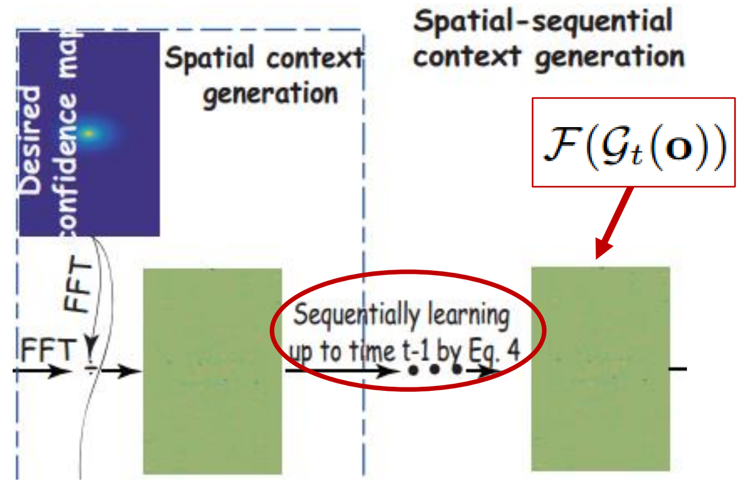


2.1.2. By κ_t^c ς_t^c

With respect to the spatial-sequential context model, the main goal is to be adaptive to estimate the translation when the target undergoes various challenging situations.

$$\boxed{\mathcal{F}(\mathcal{G}_t(\mathbf{o}))} = (1 - \xi)\mathcal{F}(\mathcal{G}_{t-1}(\mathbf{o})) + \xi\mathcal{F}(g_t(\mathbf{o})). \quad (4)$$

$$f(\mathbf{M}(I_t, r); \kappa_t^c, \varsigma_t^c) = \mathcal{G}_t(\mathbf{o}) \otimes (I_t(\mathbf{o}) w_t(\mathbf{o} - \bar{\mathbf{o}})). \quad (5)$$



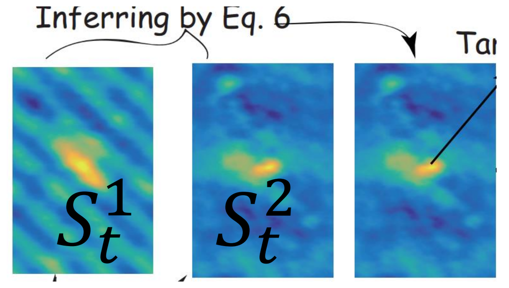
2.1.3. By κ_t^c ς_t^c ρ_t^c

In terms of the spatial-sequential-spectral context, our goal is to select the best spectral band according to the discriminative ability of κ_t^c ς_t^c in each spectra.

$$S_t^k = \mathcal{F}^{-1}(\mathcal{F}(f(\mathbf{M}(I_t^k, r); (\kappa_t^c)^k, (\varsigma_t^c)^k)))$$

$$(\rho_t^c)^k = max(S_t^k) - mean(S_t^k)$$

$$f(\mathbf{M}(I_t, r); \kappa_t^c, \varsigma_t^c, \rho_t^c) = \boxed{\arg\max_k (\rho_t^c)^k} \quad (6)$$



2.2. Fully-occlusion handling

We introduce the fast least trimmed squares (FAST-LTS) model [1] to regress the object trajectory with the tracked target centroids collected from a short previous time, and makes a trajectory prediction for upcoming frames.

$$\mathbf{u} = \textit{FAST-LTS}(\mathcal{P}_x, \mathcal{P}_y, j), \quad (7)$$

Collected object centroids (x-axis, y-axis)

Regression order(set as 1)

[1] P.. Rousseeuw and K. Driessen, "Computing lts regression for large data sets," *Data Mining and Knowledge Discovery*, vol. 12, no. 1, pp. 29–45, 2006.

2.3. Scale adaptation

This work introduces the contour closure to estimate the target scale. In particular, EdgeBox is used on color image. When the thermal infrared band is selected, we pseudo-colorize the selected frame.

$$B_t = \arg\max_{B_t{}^p} \left((\bar{B}_{t-1} \cap B_t{}^p) / (\bar{B}_{t-1} \cup B_t{}^p) \right) \lambda(B_t{}^p), \quad (8)$$

the estimated object bounding box in previous time

The object score assigned by Edgebox

Updating: $\boxed{\bar{B}_t = (1 - \eta)\bar{B}_{t-1} + \eta B_t}$

3. Experiments

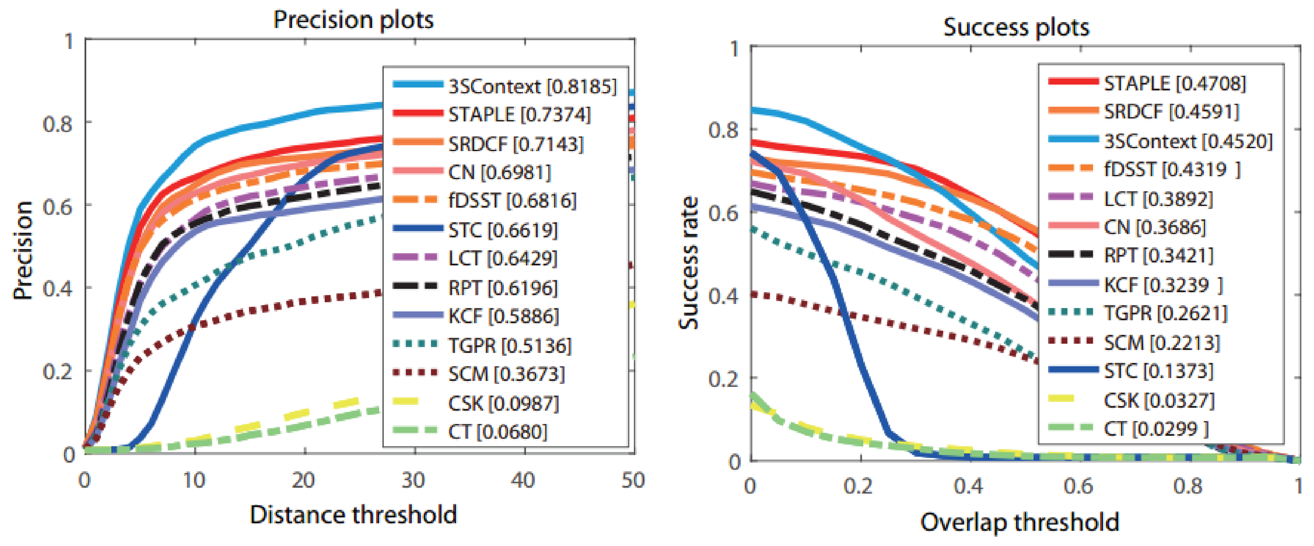


Fig. 1. The overall performance on all the sequences.

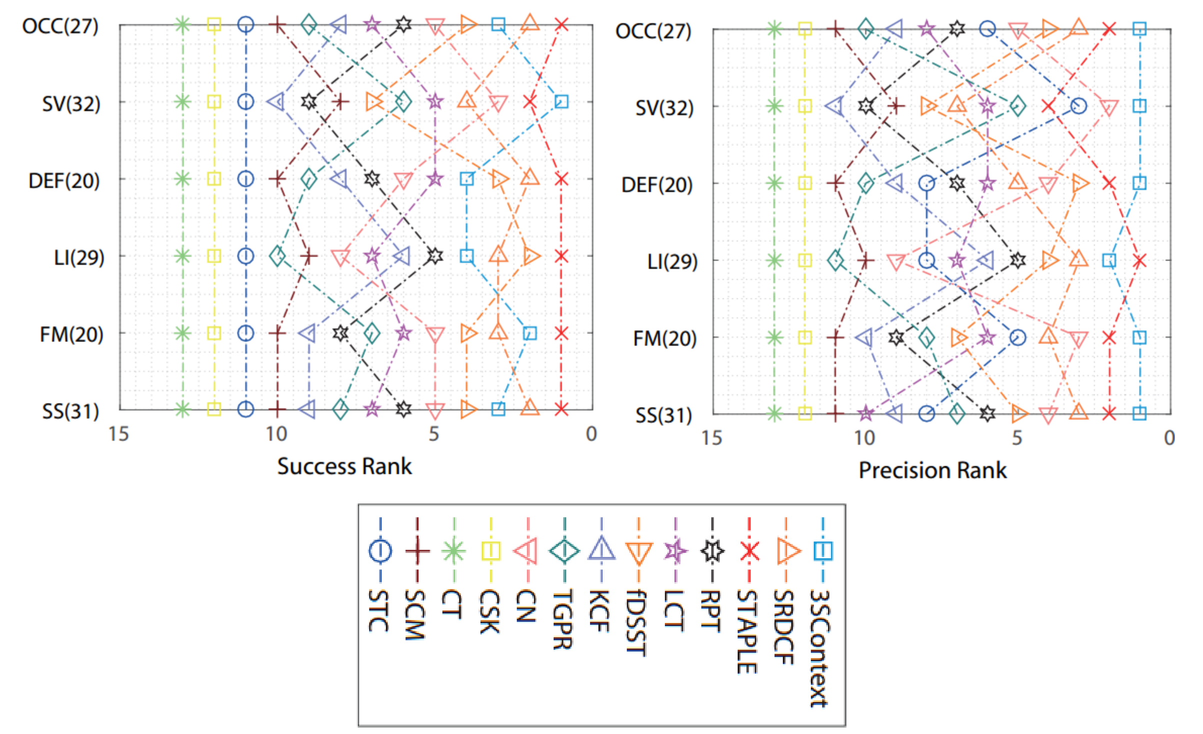


Fig. 2. The performance rank of trackers corresponding to success rate and precision on each challenging attribute. Rank-1 represents the best tracker.

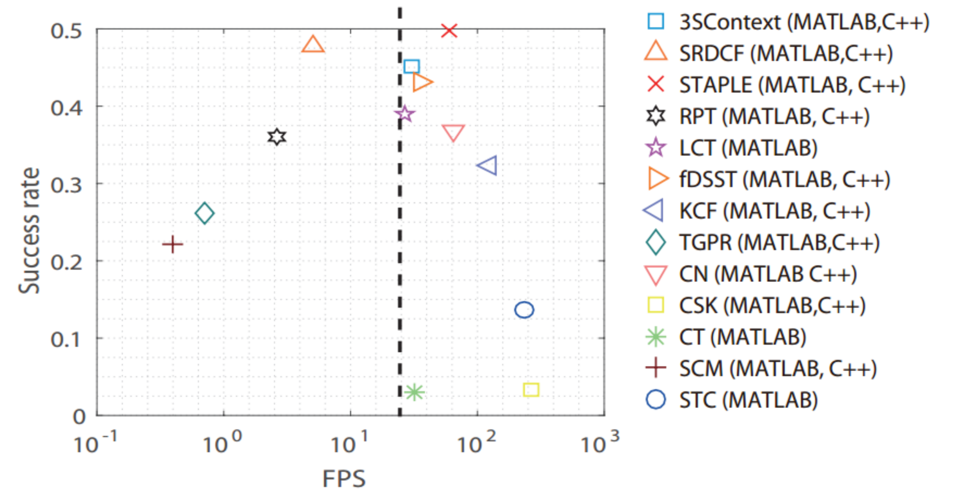


Fig. 3. Success rate vs. running speed. The x-axis is shown with a semilog display.

Remarks: We find that our spatial-sequential-spectral context can boost the precision and success rate respectively at 20% and 30% rate of STC only with spatial-temporal context exploitation. From the above analysis, we can conclude that our 3SContext demonstrates a state-of-the-art performance, especially for the precision and scale adaptation.

ICIP 2017 IEEE International Conference on Image Processing September 17-20, 2017, Beijing, China