Discriminative Correlation Filter with Channel and Spatial Reliability

Winner Publication of the VOT2017 Challenge

Leontios Mavropalias

For XXXXXXX

December 8, 2021

Overview

1. Introduction
2. Theory
3. Results
4. Conclusion

Introduction

The problem to be addressed;

- CSRT addresses the issue of *short-term tracking*.
- It may include moving deformable objects such as hands or rigid objects that change scale in the perspective; such as panoramic videos of a highway.
- Short-term tracking is challenging; factors that need to be addressed are:
- Occlusion, illumination change, fast object or camera motion, appearance changes due to rigid or non-rigid deformations and similarity to the background.
- CSRT does a good job at addressing the above issues – it won the Visual Object Tracking VOT2017 Challenge [VOT2017, 2017], which was part of the annual IEEE conference.

Introduction

It relies on *Discriminative Correlation Filters* (Discriminative – to construct a classifier or regressor to probablistically distinguish the target from its background, Correlation – template matching). CSRT relies on three high-level stages.

1. The *spatial reliability map* adjusts the filter support to the part of the object suitable for tracking. This allows the tracking of deformable objects under different scales.
2. An *optimization loop* tracks some features over time and adjusts the filter response so that a feature matching loss is minimized.
3. *Reliability scores* reflect channel-wise quality of the learned filters and are used as feature weighting coefficients in localization.

Introduction

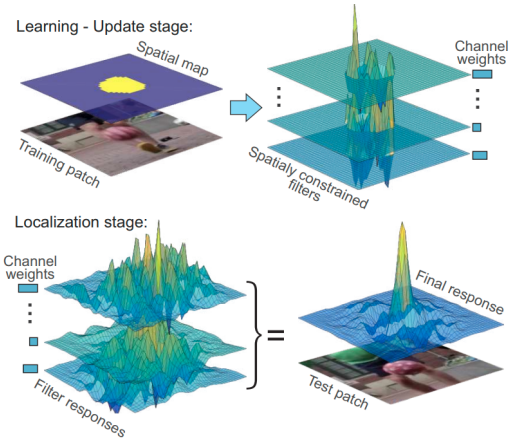


Figure: Overview of CSRT; Top left: Reliability (binary) map. Top right: Correlation filter response. Bottom left: Channel coefficient weights obtained by constrained optimisation of various correlation filters. Bottom right: Weight-averaged filter response [CSRT, 2017].

Theory; Spatial Reliability Map

Spatial reliability map $m \in [0,1]^{d_w \times d_h}$, with elements $m_{ij} \in \{0,1\}$, indicates the learning reliability of each pixel. Probability of pixel $\mathbf{x}$ being reliable conditioned on appearance $\mathbf{y}$ is specified by a Bayesian model:

$$p(m=1|\mathbf{y},\mathbf{x}) \propto \underbrace{p(\mathbf{y}|m=1,\mathbf{x})}_{p_1}\underbrace{p(\mathbf{x}|m=1)}_{p_2}\underbrace{p(m=1)}_{p_3} \quad (1)$$

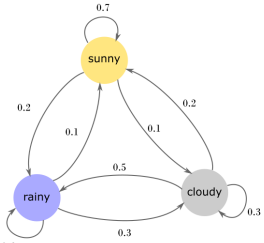
- Term p_1 (*appearance likelihood*) is computed by Bayes rule from the object foreground/background color models, which are maintained during tracking as color histograms.
- Term p_3 (*prior*) is defined as the ratio between the region sizes for foreground/background histogram extraction.

Theory; Spatial Reliability Map

- Term p_2 (*deformation invariance* of central pixels) is derived by an Epanechnikov (parabolic) kernel which favors the pixels in the center by weighting them by 0.9 and is biased against pixels away from it, weighting them down up to 0.5.

$$p_2 = p(\mathbf{x}|m = 1) = 1 - \frac{r^2}{\sigma^2} \tag{2}$$

Finally, note that the spatial reliability model takes into account the previous frames for each pixel in order to label it; "Spatial consistency of labeling m is enforced by using (1) as unary terms in a Markov random field." [CSRT, 2017]



Left: A weather Markov model.

Theory; Correlation Filter Optimization

- Correlation filter optimization (a.k.a. *filter learning*) aims to derive a filter that is iteratively correlated with the image, essentially to find the center of tracking.
- Is uses the constrained optimisation technique of Lagrangian (same technique used in mechanics to derive the minimum action path).
- Iterative techniques to minimise the Lagrangian exist and in the end the correlation filter $\mathbf{h}_c$ that minimises is it noted and recorded.

Theory; Correlation Filter Optimization

- The neat this about this technique is the although is involves convolutions in the time domain, the quantities are first transformed into the frequency domain using a Fast Fourier Transform (FFT), multiplied instead of convoluted, and then transformed back to the time domain using an Inverse FFT ($t \xrightarrow{\text{FFT}} f \xrightarrow{\text{IFFT}} t$). The complexity of both FFT and IFFT is $\mathcal{O}(n \log n)$, which makes such an optimisation extremely efficient!
- It is not only efficient, but also concise, as the authors mention that they implemented the optimisation stage in 5 lines of Matlab (next slide).

Theory; Correlation Filter Optimization

the Lagrange multiplier is updated as

$$\hat{\mathbf{l}}^{i+1} = \hat{\mathbf{l}}^{i} + \mu(\hat{\mathbf{h}}_c^{i+1} - \hat{\mathbf{h}}^{i+1}). \tag{8}$$

The minimizations in (6) have a closed-form solution:

$$\hat{\mathbf{h}}_c^{i+1} = (\hat{\mathbf{f}} \odot \overline{\hat{\mathbf{g}}} + (\mu \hat{\mathbf{h}}_m^{i} - \hat{\mathbf{l}}^{i})) \odot^{-1} (\hat{\mathbf{f}} \odot \overline{\hat{\mathbf{f}}} + \mu^{i}), \tag{9}$$

$$\mathbf{h}^{i+1} = \mathbf{m} \odot \mathcal{F}^{-1}[\hat{\mathbf{l}}^{i} + \mu^{i} \hat{\mathbf{h}}_c^{i+1}] / (\frac{\lambda}{2D} + \mu^{i}). \tag{10}$$

Algorithm 1 : Constrained filter optimization.

Require:
Image patch features $\mathbf{f}$, ideal correlation response $\mathbf{g}$, binary mask $\mathbf{m}$.

Ensure:
Optimized filter $\widehat{\mathbf{h}}$.

Procedure:
1: Initialize filter $\widehat{\mathbf{h}}^{0}$ by $\mathbf{h}_{t-1}$.
2: Initialize Lagrangian coefficients: $\widehat{\mathbf{l}}^{0} \leftarrow$ zeros.
3: **repeat**
4: Calculate $\hat{\mathbf{h}}_c^{i+1}$ from $\hat{\mathbf{h}}^{i}$ and $\hat{\mathbf{l}}^{i}$ using (9).
5: Calculate $\mathbf{h}^{i+1}$ from $\hat{\mathbf{h}}_c^{i+1}$ and $\hat{\mathbf{l}}^{i}$ using (10).
6: Update the Lagrangian $\hat{\mathbf{l}}^{i+1}$ from $\hat{\mathbf{h}}_c^{i+1}$ and $\mathbf{h}^{i+1}$ (8).
7: **until** stop condition

Figure: Filter learning optimisation implementation in the original paper [CSRT, 2017].

Theory; Channel reliability estimation

- The channel reliability at target localization stage is computed as the product of two measures; (1) a *learning channel reliability* and (2) a *detection reliability* measure.
- Measure (1) is determined by two quantities by the minimisation of the Lagrangian in the previous optimisation stage, namely $\mathbf{h}_c$ and $\mathbf{f}$:

$$\mathcal{L}(\hat{\mathbf{h}}_c, \mathbf{h}, \hat{\mathbf{l}}|\mathbf{m}) = \left|\left|\hat{\mathbf{h}}_c^H diag(\hat{\mathbf{f}}) - \hat{\mathbf{g}}\right|\right|^2 + \frac{\lambda}{2}||\mathbf{h}_m||^2 + \left[\hat{l}^H(\hat{\mathbf{h}}_c - \hat{\mathbf{h}}_m) + \overline{\hat{l}^H(\hat{\mathbf{h}}_c - \hat{\mathbf{h}}_m)} + \mu\left|\left|\hat{\mathbf{h}}_c - \hat{\mathbf{h}}_m\right|\right|^2\right] \quad (3)$$

- It is not important to remember this equation, however the quantities $\mathbf{h}_c$ and $\mathbf{f}$ are important.

Theory; Channel reliability estimation, measure 1 (learning)

- It turns out that the convolution $\mathbf{h}_c * \mathbf{f}$ defines the *channel learning reliability*. Particularly, the reliability $w_d := p(\mathbf{h}_c, \mathbf{f})$ is essentially the normalised maximum of their convolution, often called "maximum response":

$$w_d = \zeta \cdot \max(\mathbf{h}_c * \mathbf{f}), \quad s.t. \sum_d w_d = 1 \tag{4}$$

Theory; Channel reliability estimation, measure 2 (detection)

- *Per-channel detection reliability* is measured by the authors as with respect to the two major modes in the response map of $w_d := p(\mathbf{h}_c, \mathbf{f})$ (previous slide), particularly as:

$$w_d^{(det)} = 1 - \min\left(\frac{\rho_{max2}}{\rho_{max1}}, \frac{1}{2}\right) \tag{5}$$

- Therefore is ρ_{max1} is large compared to ρ_{max1}, then the detection (object detection) reliability of the said channel is good.
- Objects moving fast in front of the camera, blocking the view, lower the detection reliability.

Theory; Channel reliability estimation, measure 2 (detection)

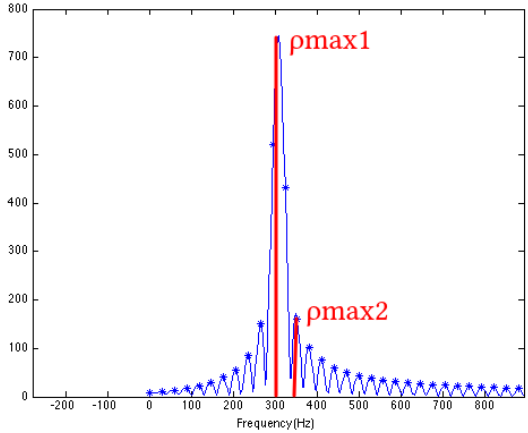


Figure: Reliability response w_d of a channel (made-up data). The particular response implies a good detection measure as ρ_{max1} is significantly larger than ρ_{max1}.

Theory; algorithm overview

The previous stages are applied to iteratively localise an object, followed by updating the reliability map (and the background/foreground model). In summary:

- *Localize*:
  1. Detection reliability $\times$ learning reliability measures = channel reliability
  2. Location = $(\sum \mathbf{h}_i) \times$ channel reliability scores, where $\mathbf{h}_i$ are the correlation filters learned at the optimization stage. the estimated channel reliability scores.

.

- *Update*.
  1. Find foreground/background histogram by a regressive scheme.
  2. The spatial reliability map is constructed in a Bayesian probabilistic way.
  3. Learn the optimal correlation filter by Lagrangian optimization.
  4. Estimate per-channel reliability.

The algorithm in its full glory is summarized in the following slide.

Theory; algorithm overview

Algorithm 2 : The CSR-DCF tracking algorithm.

Require:

Image $\mathbf{I}_t$, object position on previous frame $\mathbf{p}_{t-1}$, scale s_{t-1}, filter $\mathbf{h}_{t-1}$, color histograms $\mathbf{c}_{t-1}$, channel reliability $\mathbf{w}_{t-1}$.

Ensure:

Position $\mathbf{p}_t$, scale s_t and updated models.

Localization and scale estimation:

1: New target location $\mathbf{p}_t$: position of the maximum in correlation between $\mathbf{h}_{t-1}$ and image patch features $\mathbf{f}$ extracted on position $\mathbf{p}_{t-1}$ and weighted by the channel reliability scores (Section 3.3).

2: Using location $\mathbf{p}_t$, estimate new scale s_t.

Update:

3: Extract foreground and background histograms $\tilde{\mathbf{c}}^f$, $\tilde{\mathbf{c}}^b$.

4: Update foreground and background histograms $\mathbf{c}_t^f = (1-\eta_c)\mathbf{c}_{t-1}^f + \eta_c\tilde{\mathbf{c}}^f$, $\mathbf{c}_t^b = (1-\eta_c)\mathbf{c}_{t-1}^b + \eta_c\tilde{\mathbf{c}}^b$.

5: Estimate reliability map $\mathbf{m}$ (Section 3.1).

6: Estimate a new filter $\tilde{\mathbf{h}}$ using $\mathbf{m}$ (Algorithm 1).

7: Estimate channel reliability $\tilde{\mathbf{w}}$ from $\tilde{\mathbf{h}}$ (Section 3.3).

8: Update filter $\mathbf{h}_t = (1-\eta)\mathbf{h}_{t-1} + \eta\tilde{\mathbf{h}}$.

9: Update channel reliability $\mathbf{w}_t = (1-\eta)\mathbf{w}_{t-1} + \eta\tilde{\mathbf{w}}$.

Results; Benchmarking definitions

- The authors used the OTB100 benchmark.
- It contains results of 29 trackers evaluated on 100 sequences by a no-reset evaluation protocol.
- The KPIs are (1) the *success*; portion of frames with the overlap (Intersection of Union - IoU) between predicted and ground truth bounding box greater than a threshold with respect to all threshold and (2) the *precision*; statistics on the object localization center error.
- In summary, CSRT is among the top two best performers, if not the best Its success is slightly lower than SRDCF, however its precision is by far the best of all.

Results; Benchmarking results

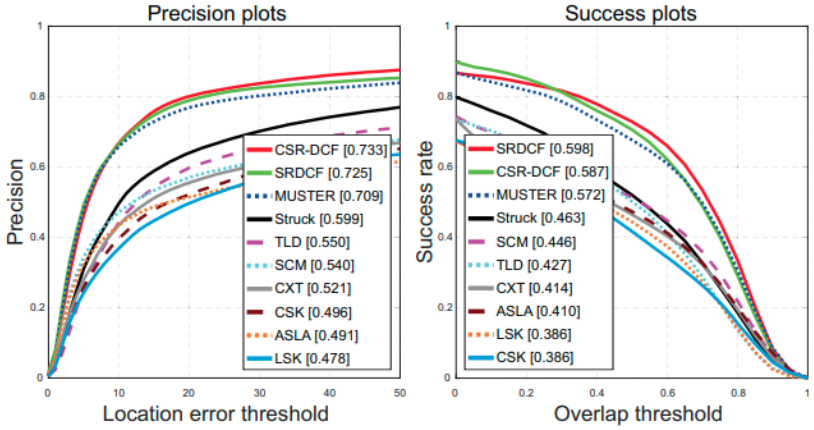


Figure: Benchmarking success results.

Results; Benchmarking results

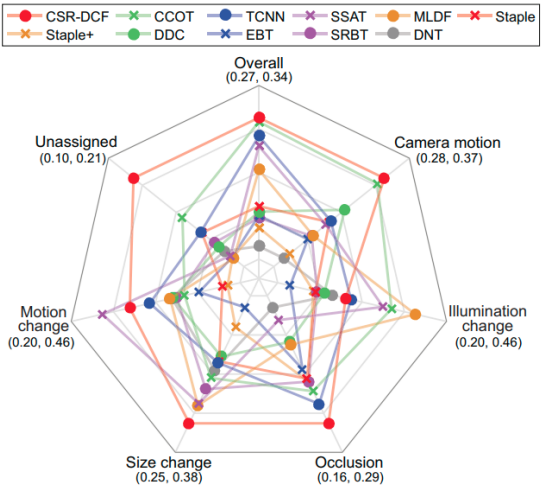


Figure: Benchmarking precision results.

Extensions; Vehicle Tracking

- CSRT can easily be integrated with high-quality object detectors such as YOLO in order to track vehicles, either a single or multiple vehicles at the same time.
- Amitha and Narayanan have built such a system for vechicle classification and counting. They reported that the accuracy of CSRT was 100% [Amitha, 2021].

Extensions; Vehicle Tracking

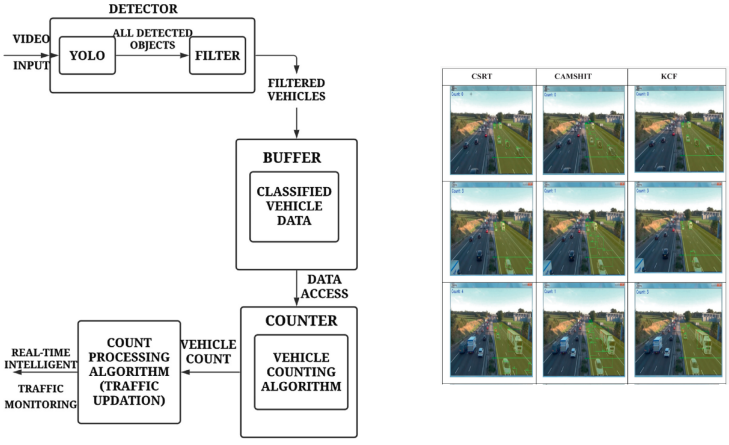


Figure: Amitha and Narayanan's [Amitha, 2021] benchmarking configuration and some results.

Conclusion

- CSRT is a state-of-the-art short-term tracking method and the top performer of 2017.
- It successfully addresses complex issues, such as deformable objects, rapid movements, and illumination changes. It is, however, lightweight, and works in real-time.
- Not very well known yet. No one at my company knew about it!
- Already implemented in the latest OpenCV. Some open-source implementations exist too.

References

M. Kristan et al (2017)
The Visual Object Tracking VOT2017 Challenge Results
2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017, 1949 – 1972.

MA. Lukezic, T. Vojir, L. C. Zajc, J. Matas and M. Krista (2017)
Discriminative Correlation Filter with Channel and Spatial Reliability
IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, 4847 – 4856.

I. Amitha, N. Narayanan (2021)
Improved Vehicle Detection and Tracking Using YOLO and CSRT
Communication and Intelligent Systems, Springer Singapore, 435 – 446.