Evaluating Template Rescaling in Short-Term Single-Object Tracking

Jörgen Ahlberg, Amanda Berg

Jörgen Ahlberg, Amanda Berg, Evaluating Template Rescaling in Short-Term Single-Object Tracking, 2015, 17th IEEE International Workshop on Performance Evaluation of Tracking and Surveillance (PETS).

Evaluating Template Rescaling in Short-Term Single-Object Tracking

Jörgen Ahlberg[1,2], Amanda Berg[1,2]
[1]Termisk Systemteknik AB, Diskettgatan 11 B, 583 35 Linköping, Sweden
[2]Computer Vision Laboratory, Dept. EE, Linköping University, 581 83 Linköping, Sweden
{jorgen.ahl,amanda.}berg@termisk.se, {jorgen.ahl, amanda.}berg@liu.se

Abstract

In recent years, short-term single-object tracking has emerged has a popular research topic, as it constitutes the core of more general tracking systems. Many such tracking methods are based on matching a part of the image with a template that is learnt online and represented by, for example, a correlation filter or a distribution field. In order for such a tracker to be able to not only find the position, but also the scale, of the tracked object in the next frame, some kind of scale estimation step is needed. This step is sometimes separate from the position estimation step, but is nevertheless jointly evaluated in de facto benchmarks. However, for practical as well as scientific reasons, the scale estimation step should be evaluated separately – for example, there might in certain situations be other methods more suitable for the task.

In this paper, we describe an evaluation method for scale estimation in template-based short-term single-object tracking, and evaluate two state-of-the-art tracking methods where estimation of scale and position are separable.

1. Introduction

Tracking of objects in video is a problem that has been and still is subject to extensive research [8]. Indicators of the popularity of the topic are challenges/contest like the recurring Visual Object Tracking (VOT) challenge [5, 6], the Linköping Thermal IR tracking benchmark (LTIR) [1], the Online Object Tracking (OTB) benchmark [8], and the series of workshops on Performance Evaluation of Tracking and Surveillance (PETS) [9].

Short-term single-object (STSO) tracking is a subproblem in object tracking that has received much attention recently. Many STSO tracking methods rely on some form of template matching, the trivial example being a sliding window and normalized cross-correlation. In order to track not only the 2D position of the object, but also a scale change, the template (or the image) needs to be resampled and matched in multiple scales. Such matching becomes computationally expensive, and other methods have been developed. With ”scale change” we refer to changes both due to the object being deformable and due to the object moving closer to or away from the camera.

Some trackers that do not have an inherent scaling capability have later been complemented with a scale estimation method, improving the tracking results in terms of accuracy and/or robustness. However, the scaling capability itself has not been evaluated. The relevance of doing that comes with the insight that at least one of the causes of scale change (camera-object distance change) can be estimated in other ways, for example by a complementary sensor (another camera observing the object from another angle or a distance-measuring sensor) or by assuming that the object moves on a known surface. The latter is the typical case for surveillance cameras; the camera is mounted several meters above ground, observing a quite flat area. In that scenario, the image coordinates of the feet of a pedestrian together with camera calibration readily reveals the distance to the pedestrian.

So why is this interesting? First, it is the application viewpoint. Evaluating the scale estimation method and comparing to other methods would give the answer to which scale estimation method should be used in a practical setting. For example, is it better to estimate the scale change from the change of position than to use the tracker's bulitin functionality? Second, it serves as an indicator whether more research is needed or not.

2. Related work

There is a plethora of STSO trackers in the literature. The most recent collection is, as far as the authors are aware, the proceedings from the VOT challenge 2014 [6]. Trackers that are useful for this experiment must fulfill the following criteria:

- They should be reasonably modern.
- There should be two variants of the tracker; one with and one without the scaling capability.

978-1-4673-7632-7/15/$31.00 ©2015 IEEE

- We should have access to the trackers, i.e. have access to the code or an executable.

Relevant trackers include DFT [7], EDFT [4], MOSSE [2] and DSST [3]:

- EDFT is an extension of the distribution field tracker (DFT), replacing the histograms with a channel coded representation. A so far unpublished extension of EDFT adds scale capability.
- DSST is an extension of the MOSSE tracker adding scale estimation and some minor improvements. We will compare the DSST tracker with a scale-disabled version, i.e., a slighly improved MOSSE. We call the scale-free tracker DSST-SF.

3. Evaluation method

In order to evaluate the scaling capability of the tracker we need at least one image sequence with an annotated ground truth in terms of a bounding box around the tracked object in each frame. The tracked object needs to have a variation in scale, for example due to varying distance to the camera. In fact, this is more interesting than scale change due to deformation, as there are alternative methods to measure the object-camera distance, and thus the scale change.

From the annotated ground truth (which arguably contains some noise), we extract the scale of the object as the height, in pixels, of the bounding box in each frame (s_{gt}^k), where k is the frame number.

Second, we apply the trackers $t1, t2, ...$ and their scale-free variants $\tau 1, \tau 2, ...$ to the sequence. The scales ($\hat{s}_1^k$, $\hat{s}_2^k$, ...) are extracted from the tracking results in the same way as from the ground truth. The relative errors are computed and averaged over the sequence, i.e.

$$e_n^k = \frac{\left|s_{gt}^k - \hat{s}_n^k\right|}{s_{gt}^k} \qquad (1)$$

$$e_n = \frac{1}{K}\sum_{k=1}^{K} e_n^k \qquad (2)$$

where K is the number of frames in the sequence and n is the number of the tracker.

Third, the scale is also estimated as a linear function of the vertical position of the lower border of the bounding box from *the scale-free version* of each evaluated tracker, i.e.,

$$z_n^k = f(y_{\tau n}^k), \qquad (3)$$

where $y_{\tau n}^k$ is the mentioned vertical position as given by the scale-free version n:th tracker. In analogy with (2), the error measure ϵ_n is computed as well.

Fourth, as a reference, we also estimate the scale as a linear function of the vertical position of the lower edge of the bounding box ($s_f^k = f(y_{gt}^k)$) and the corresponding error ϵ_0. If the object is completely rigid, s_f should equal s_{gt}. ϵ_0 serves a lower bound for ϵ_n.

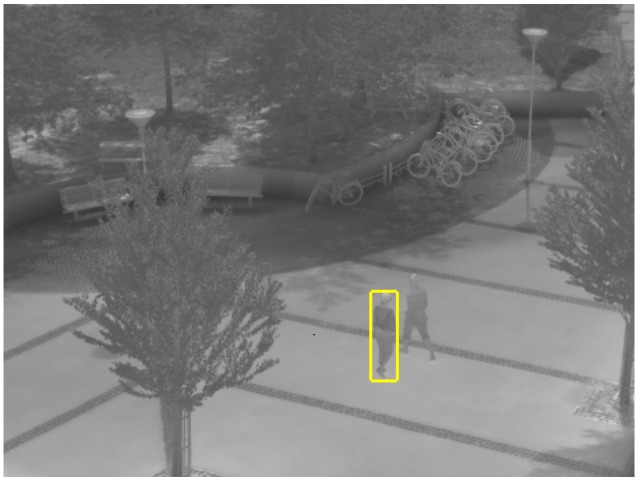

Figure 1: One frame from the *depthwise crossing* sequence in the LTIR dataset. The annotated bounding box is marked in yellow.

The first interesting result is the errors e_n as they give a quantitative way of comparing different ways of estimating the object scale and a (usually subjective) measure of if the scale estimate is good enough.

The second interesting result is the relation between e_n and ϵ_n. If the scale error e_n from the trackers scale estimate is larger than the estimate ϵ_n made from the object position, the latter method is preferable in a practical application (if applicable, that is).

4. Experiment

4.1. Input data

We have chosen sequence #15 from the publicly available LTIR[1] dataset [1], called *depthwise crossing*. The tracked object in the sequence is a person walking on a flat surface (the ground), for most of the sequence in the direction away from the (stationary) camera but turning towards the camera close to the end of the sequence. The camera is a thermal infrared one, but this is of no importance for this experiment - it could as well have been a ordinary visual light video camera.

4.2. Trackers

We have applied four trackers to the selected sequence: DSST; DSST-SF; a modified EDFT (here called EDFT-Th since it is adapted to thermal imagery); and a scale-capable

[1] Dataset available at www.cvl.isy.liu.se/research/datasets/ltir. This is the dataset used for the VOT-TIR challenge 2015, www.votchallenge.net.

Table 1: Errors

Method	Mean absolute error [pixels]	Mean relative error [%]
Position/GT	1.2	$\epsilon_0 = 1.6$
Position/DSST-SF	3.8	$\epsilon_2 = 5.2$
Position/EDFT-Th	2.0	$\epsilon_1 = 2.7$
DSST	11.5	$e_2 = 15.7$
EDFT-S	6.4	$e_1 = 7.7$

of EDFT-Th variant called EDFT-S. The exact workings of EDFT-Th and EDFT-S are out of scope for this paper and will be published elsewhere.

4.3. Results

The scale, that is, the height in pixels, of the tracked object is plotted in Figure 2. The dashed lines show the estimates using object positions and the scale-free trackers. The relative errors are shown in Figure 2 and given in Table 1. To give an idea about how large (or small) the errors are in the image plane, the mean absolute errors (in pixels) are given in the table as well.

For the EDFT-S tracker, the scale estimate is worse when using position/scale-free tracker in the first half of the sequence, but actually they behave similarly in the second half. On average, the error is around three times as large, and exploiting the position in a fixed-camera-known-ground scenario is thus still a good idea.

For the DSST tracker, the results are somewhat ambiguous (and more interesting). In the beginning of the sequence, the estimate using position/scale-free tracker is better, but as the object becomes smaller, this estimate gets worse, until the DSST-SF tracker fails and must be reinitialized from the ground truth. This also makes the estimate using the position/scale-free tracker better in the rest of the sequence. This is particularly interesting since it points out that scale estimate (by DSST) does not only give a better approximation of the bounding box and thus improving the accuracy, but also improves the trackers' robustness (compared to DSST-SF). Using the scale-free version, the tracker is more likely to fail when the object becomes smaller and the tracked bounding box does not. The main reason for this is that more and more background will be introduced in the model.[2]

As a note, the scale estimation methods of the DSST tracker seemed inferior to the scale estimation method of the EDFT-S tracker. This is somewhat misleading, since we made the evaluation on a thermal infrared image sequence, which the EDFT-S is specially designed for.

5. Conclusion

The proposed evaluation method clearly shows the difference between the evaluated trackers' scale estimation method and estimation from position only. The results also indicate that there is room for improvement of the particular two trackers evaluated here. Estimating scale from object position is of course only relevant under certain conditions (such as known surface, known camera, non-flying objects), but is still relevant since these conditions are quite commonly fulfilled and since it can serve as a goal to reach for tracking method developers.

For both trackers, better results are reached by using the scale free variation and then estimating the scale change from the tracked position. If this scale change was actually inferred to the tracker in each frame, during tracking, it is quite safe to assume that the tracking performance (in terms of accuracy and robustness) would improve. In conclusion, the state-of-the-art trackers would not always be preferable compared to a less advanced version.

As a future work, we will modify the evaluated scale-free versions of the trackers so that they take the scale estimated by the position into account, in each frame, before continuing the tracking. This will most likely improve the performance of the trackers, and since this is rthe way a real system would be implemented, the comparison would be more relevant.

Acknowledgment

The authors gratefully acknowledges the financial support from the Swedish Research Council through the project ”Learning Systems for Remote Thermography”, grant no. D0570301, as well as the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 312784 (P5). Moreover, we thank M. Danelljan and M. Felsberg at Linköping University for providing the source code of the DSST and EDFT trackers, respecively.

References

[1] A. Berg, J. Ahlberg, and M. Felsberg. A thermal object tracking benchmark. In *Proc. 12th IEEE Int. Conf. on Advanced Video and Signal based Surveillance (AVSS)*, August 2015.

[2] D. Bolme, J. Beveridge, B. Draper, and Y. M. Lui. Visual object tracking using adaptive correlation filters. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[3] M. Danelljan, G. Häger, F. Shahbaz Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *Proc. of the British Machine Vision Conf. (BMVC)*, 2014.

[2] The EDFT-Th tracker includes a mechanism for mitigating background contamination in the model, and is thus less sensitive.

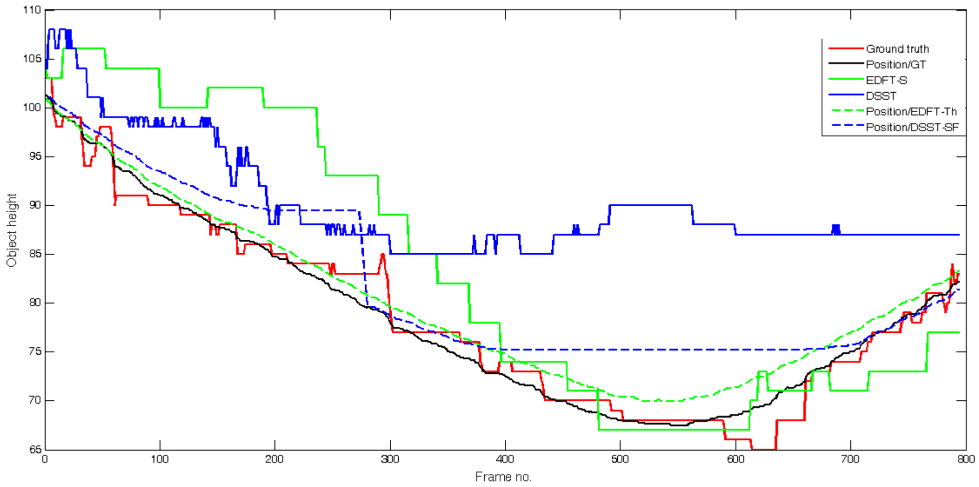

Figure 2: Estimated scales. Note that the *red* line shows the ground truth, not the black one. The sudden jump to in the blue dashed line around frame 280 is due to the DSST-SF tracker losing track and being reinitialized.

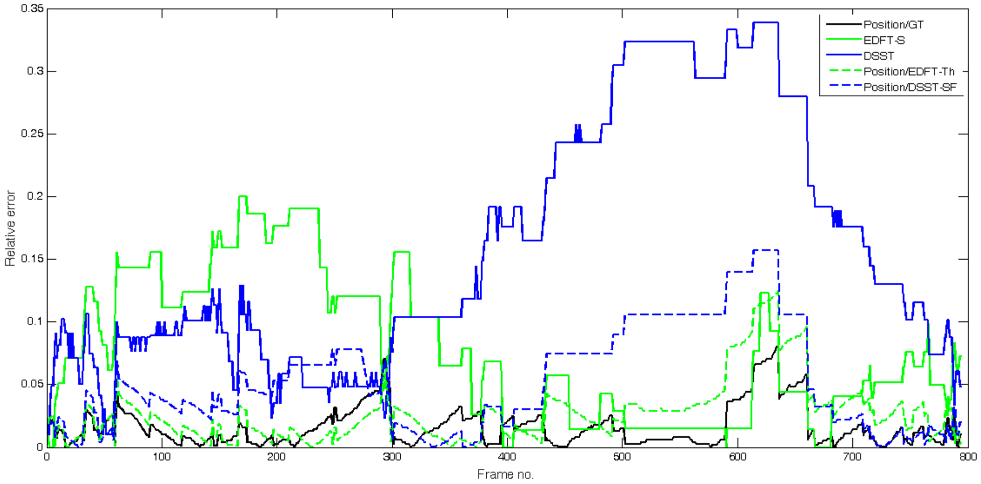

Figure 3: Relative errors.

[4] M. Felsberg. Enhanced Distribution Field Tracking using Channel Representations. In *Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, 2013.

[5] M. Kristan et al. The visual object tracking VOT2013 challenge results. In *Proc. IEEE Int. Conf. on Computer Vision Workshops (ICCVW)*, 2013.

[6] M. Kristan et al. The Visual Object Tracking VOT2014 challenge results. In *Workshop on Visual Object Tracking Challenge (VOT) - ECCV workshop*, LNCS, Springer, 2014.

[7] L. Sevilla-Lara and E. G. Learned-Miller. Distribution fields for tracking. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012.

[8] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2013.

[9] D. P. Young and J. M. Ferryman. PETS metrics: Online performance evaluation service. In *Proc. 14th Int. Conf. Computer Communications and Networks (ICCCN)*, 2005.