Real-time part-based visual tracking via adaptive correlation filters

Ting Liu[1], Gang Wang[1], Qingxiong Yang[2]
[1]School of Electrical and Electronic Engineering, Nanyang Technological University. [2]Department of Computer Science,City University of Hong Kong.

Correlation filters have been used in tracking tasks recently because of the high efficiency [1, 2]. However, the conventional correlation filter based trackers cannot deal with occlusion. In this paper, we propose a novel tracking method which track objects based on parts with multiple correlation filters. Extensive experiments have been done to prove the effectiveness of our method.

Our key idea is to adopt the correlation filters as part classifiers. Hence, the part evaluation speed can be fast. Our contribution is to develop new criteria to measure the performance of different parts, and assign proper weights to them. Specifically, we propose to use Smooth Constraint of Confidence Maps as a criterion to measure how likely a part is occluded. Besides, we developed the spatial layout constraint method to: 1) effectively suppress the noise caused by combining of individual parts, 2) estimate the correct size of bounding box when the target is occluded.

The location of the target object is given in the 1st frame of the video, and the tracker is then required to track the object (by predicting a bounding box containing the object) from the 2nd frame to the end of the video. The target is divided into several parts. For each part of the object we run an independent KCF tracker [2] that outputs a response map. The maps are then combined to form a single confidence map for the whole target that is used in the Bayesian inference framework [3, 4].

For correlation filter based classifier, the peak-to-sidelobe ratio (PSR) (Eq. 3) can be used to quantify the sharpness of the correlation peak. In addition, for tracking problems, the temporal smoothness property is helpful for detecting whether the target is occluded. Taking this observation into consideration and as validated from our experiments, we propose that the smooth constraint of confidences maps should be considered for the weight parameters. We define the smooth constraint of confidence maps (SCCM) in Eq. 4. The joint confidence map at the t-th frame is defined as:

$$C^t = \sum_{i=1}^{N} \mathrm{w}_i^t \hat{f}_{\mathrm{p}(i)}^t, \tag{1}$$

where $\hat{f}_{\mathrm{p}(i)}^t$ is the confidence map of the i-th part at time t. $\mathrm{p}(i)$ denotes the relative position of part response in the joint confidence map C^t; it is determined by the maximum value of the part confidence map. N is the number of parts used to divide the target. w_i^t is the weight parameter of corresponding part.

$$\mathrm{w}_i^t = PSR_i + \eta \cdot \frac{1}{SCCM_i} \tag{2}$$

$$PSR_i = \frac{\max\left(\hat{f}_{\mathrm{p}(i)}^t\right) - \mu_i}{\sigma_i}. \tag{3}$$

$$SCCM_i = \left\| \hat{f}_{\mathrm{p}(i)}^t - \hat{f}_{\mathrm{p}(i)}^{t-1} \oplus \Delta \right\|_2^2 \tag{4}$$

where, μ_i and σ_i are the mean and the standard deviation of the i-th confidence map respectively. η is the trade-off between correlation sharpness and smoothness of confidence maps; in our experiments, it is simply set as 1. $\oplus$ means a shift operation of the confidence map, and Δ denotes the corresponding shift of maximum value in confidence maps from frame $t-1$ to t. $\hat{f}_{\mathrm{p}(i)}^{t-1}$ and $\hat{f}_{\mathrm{p}(i)}^t$ denote the individual response maps of part i.

A threshold is used to adaptively update each part tracker separately. The learning rate for the model is set proportional to the weight value (Eq. 2).

$$\mathcal{F}(\alpha)_i^t = \begin{cases} (1-\beta \mathrm{w}_i^t)\mathcal{F}(\alpha)_i^{t-1} + \beta \mathrm{w}_i^t \mathcal{F}(\alpha)_i & if\ \mathrm{w}_i^t > \text{threshold} \\ \mathcal{F}(\alpha)_i^{t-1} & \text{else} \end{cases} \tag{5}$$

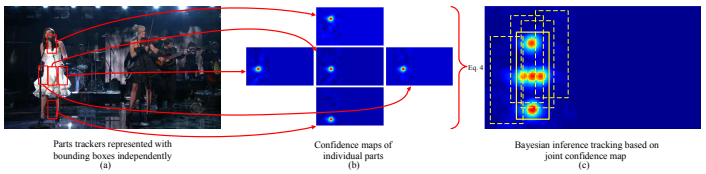

Figure 1: Each part tracker independently tracks the corresponding part and outputs a response map. The separate response maps are combined by Eq. 1. We track the whole target based on the joint confidence map in the Bayesian inference framework.

$$\hat{x}_i^t = \begin{cases} (1-\beta \mathrm{w}_i^t)\hat{x}_i^{t-1} + \beta \mathrm{w}_i^t x_i & if\ \mathrm{w}_i^t > \text{threshold} \\ \hat{x}_i^{t-1} & \text{else} \end{cases} \tag{6}$$

we carry out the tracking problem as a Bayesian inference task. Let s^t denote the state variable describing the affine motion parameters of an object at the time t (e.g. location or motion parameters) and define $O^t = [o^1, o^2, \cdots, o^t]$ as a set of observations with respect to joint confidence maps. The optimal state $\hat{s}^t$ is computed by the maximum a posterior (MAP) estimation

$$\hat{s}^t = \arg\max_{s_j^t} p(s_j^t \,|\, O^t) \tag{7}$$

where s_j^t is the state of the j-th sample. The observation model $p(o^t|s^t)$ denotes the likelihood of the observation o^t at state s^t. Maximizing the posterior in Eq. 7 is equivalent to maximizing the likelihood $p(o^t|s^t)$.

We introduce a spatial layout constraint mask to enforce the spatial constraint. Hence, the observation model is constructed by

$$p(o^t|s^t) = \frac{1}{|M^t|} \sum C^t(s^t) \odot M^t \tag{8}$$

where $\odot$ is the element-wise production. M^t is a spatial layout constraint mask built by N cosine windows which gradually reduce the pixel values near the edge to zero. The relative positions of these cosine windows are determined by the maximum value of corresponding part response maps. The size of cosine window is determined by the size of tracking part. $|M^t|$ is the number of pixels within the mask. $C^t(s^t)$ means a candidate sample.

From the experiment results we can see that our method performs better or at least comparable with the other state-of-the-art trackers. Our conclusion is that by using the adaptive weighting, updating and structural masking methods, our tracker is robust to occlusion, scale and appearance changes.

[1] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. Visual object tracking using adaptive correlation filters. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2544–2550. IEEE, 2010.

[2] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2015. doi: 10.1109/TPAMI.2014.2345390.

[3] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *International Journal of Computer Vision*, 77(1-3):125–141, 2008.

[4] Dong Wang, Huchuan Lu, and Ming-Hsuan Yang. Least soft-threshold squares tracking. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2371–2378. IEEE, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.